Universidade de Lisboa
Faculdade de Ciências
Departamento de Estatística e Investigação Operacional

UNIVERSIDADE
DE LISBOA

Aplicação dos Modelos Lineares Generalizados às Telecomunicações Móveis: caracterização dos clientes que desactivam os seus serviços

Paula Figueiredo Mestre

Mestrado em Probabilidades e Estatística

2009

Universidade de Lisboa
Faculdade de Ciências
Departamento de Estatística e Investigação Operacional

UNIVERSIDADE
DE LISBOA

Aplicação dos Modelos Lineares Generalizados às Telecomunicações Móveis: caracterização dos clientes que desactivam os seus serviços

Dissertação Orientada pela
Professora Doutora Teresa Alpuim

Paula Figueiredo Mestre

Mestrado em Probabilidades e Estatística

2009

Resumo

O mercado de telecomunicações é actualmente caracterizado por forte concorrência entre os vários operadores em actividade e por um elevado nível de saturação, sendo cada vez mais difícil a angariação de novos clientes, havendo por isso uma forte aposta na retenção e fidelização dos existentes.

Por este motivo é cada vez mais importante o conhecimento do perfil do cliente, a identificação dos factores que influenciam a sua satisfação e das variáveis que o influenciam realmente na decisão de mudar de prestador de serviço ou de se manter com o actual.

Com o objectivo de caracterizar os clientes que desactivam os seus serviços recolheu-se informação relativa a um segmento específico: clientes residenciais pós-pagos. A formulação de um (ou vários) modelo(s) de regressão logística - para variável resposta do tipo binário, activo ou desactivo – servirá para identificar quais são os factores que estes clientes mais valorizam e têm real impacto na sua satisfação, bem como, em oposição, identificar claramente quais os podem levar à decisão de mudar de operador, ou seja, encontrar os factores que diferenciam os clientes activos dos desactivos.

Os modelos de regressão logística são um caso particular de um vasto conjunto de modelos de utilização muito ampla: os modelos lineares generalizados. Estes caracterizam-se pelo facto de poderem ter variável resposta não normal, desde que esta satisfaça a condição de ser bem ajustada por uma distribuição pertencente à família exponencial. A ligação entre a variável resposta e o vector de covariáveis pode ser estabelecida através de uma função monótona diferenciável chamada função de ligação. É apresentado neste estudo um método de estimação dos parâmetros para este tipo de modelos. São descritas várias estratégias de modelação e comparados os resultados respectivos, sendo também descritos alguns problemas numéricos surgidos durante o processo (comuns para dados deste tipo), algumas possíveis causas e soluções.

Palavras Chave:
Telecomunicações, Modelos Lineares Generalizados, Regressão Logística, Resposta Binária

Abstract

Telecomunications business is presently marked by fierce competition amongst operators and high saturation level, therefore leading to growing difficulties to acquire new customers. Due to this situation market players are increasingly focusing on retention and loyalty programs to maintain current ones.

Strong knowlegde of customer profile is gaining great importance, since knowing what are the factors that influence customers' satisfaction and can make one decide to change service provider (or keep the present one) can be of great help to design retention programs and focus on decisive / really important variables.

With the goal of characterizing deactivated customer profile, all available information related with a specific segment – post-paid residentials – has been gathered. Regression models – for binary dependent variables, active or deactivated – were formulated based on this data. These models are aimed to help identify which factores are valued and have a real impact on customers, and to find out which can lead them to the decision of changing to another provider. The purpose is therefore to identify the factors which differentiate active customers form deactivated ones.

Logistic regression models are a particular case of a much wider class of models vastly used: generalized linear models. These can have a non linear response variable, as long as it is well aproximated by any distribution belonging to the exponential family. The relationship between the dependent variable and the independent ones can be established through a differentiable and monotone function called the link function. An estimation method for the model parameters is presented in this paper.

Disctinct modelling strategies are described and compared in this study. Some numerical problems (common for this type of data) arisen during the modelling process are also detailed, as well as their possible causes and some solutions.

Key Words:
Telecomunications, Generalized Linear Models, Logistic Regression, Binary Response

Índice

1 – Introdução

Na generalidade dos países europeus o mercado de telecomunicações móveis é neste momento caracterizado por uma forte saturação por terem sido atingidas taxas de penetração elevadíssimas. Portugal não é excepção, tendo esta taxa um valor muito próximo dos 100%, criando consequentemente grande dificuldade na angariação de novos clientes.

Este mercado caracteriza-se por uma grande variedade de produtos e serviços, com uma forte concorrência quer ao nível da oferta disponibilizada aos clientes (novos produtos, serviços, etc..) quer ao nível dos preços (tarifários mais competitivos). Tendo em conta o tipo de serviço prestado é relativamente fácil para os clientes mudarem de prestador, registando-se por esse motivo uma elevada mobilidade entre operadores.

Face a este cenário, as operadoras de telecomunicações têm de apostar na diferenciação como forma de reter e fidelizar os seus clientes. Dada a complexidade do serviço prestado e de toda a infra-estrutura que o suporta, esta diferenciação baseia-se num grande conjunto de factores, dos quais se podem destacar os seguintes:

- Qualidade do serviço prestado
  O serviço prestado tem de ter uma elevada qualidade em todas as suas vertentes, quer ao nível da utilização, quer ao nível do suporte (esclarecimento de dúvidas, interacções do cliente com os serviços de atendimento, utilização de ferramentas de self-service, etc…)

- Fiabilidade do serviço
  A fiabilidade está interligada com a qualidade, o cliente valoriza um serviço sem falhas, por exemplo ao nível da rede, dos serviços de apoio ao cliente, facturação rigorosa, clara e isenta de erros, etc…

- Simplicidade
  A utilização dos serviços disponíveis deve ser simples e intuitiva.

- Preço
  Os serviços prestados devem ter um preço que o cliente considere adequado e justo (havendo também que ter em conta as questões concorrenciais já mencionadas).

- Inovação
  Existe uma forte aposta no desenvolvimento de novos produtos, mais próximos das necessidades e expectativas dos clientes, baseados em novas tecnologias.
  Tem havido nos últimos anos uma enorme evolução nas tecnologias associadas às telecomunicações, o que obriga os prestadores a grande dinamismo para garantir que disponibilizam aos seus clientes os serviços mais avançados, o que de melhor existe no momento.

- Diversidade
  O mercado de consumidores caracteriza-se por uma grande variedade ao nível do tipo de serviço pretendido: os clientes empresariais têm características completamente distintas dos clientes residenciais, e dentro de cada um destes grupos há necessidades muito diversas.
  Há clientes que preferem serviços pré-pagos, outros pós-pagos, podem valorizar ou não comunicação por SMS e MMS, a utilização de Roaming, ter ou não tráfego de dados, utilizar ou não serviços de self-service (por exemplo baseados na internet), ter ou não equipamentos mais sofisticados que permitam utilização de eMail, Internet, etc…
  Assim, devem ser disponibilizados serviços adequados às expectativas dos clientes, o que obriga a grande diversidade a todos os níveis: tarifários, equipamentos, tipo de serviço, etc…

- Convergência

Associada à simplicidade, os clientes valorizam a convergência: poderem utilizar os serviços disponíveis através de vários canais ou equipamentos (por exemplo o telefone e o computador) e em qualquer localização, poderem ter serviços de mais que um tipo num único equipamento (serviços comunicações, internet, acesso ao eMail, aplicações, etc…), terem uma factura única com todos os serviços prestados, etc…

Muitos outros factores poderiam ser referidos e aprofundados, este resumo pretende apenas aqui mostrar a complexidade associada à oferta de telecomunicações e à elevada expectativa / exigência dos clientes com este tipo de serviço.
Os prestadores devem investir no estabelecimento de uma relação de confiança com o cliente, no aumento dos seus níveis de satisfação e consequentemente na sua fidelização. Para isso deve haver uma forte aposta na retenção e no conhecimento do cliente. Havendo um tão vasto conjunto de factores que o podem influenciar, é imprescindível saber quais são os que mais valoriza e têm real impacto na sua satisfação, bem como no lado oposto, identificar claramente quais o podem levar à decisão de mudar de operador.
A identificação destes factores é evidentemente uma enorme mais-valia, pois permite traçar planos de melhoria nos processos realmente valorizados pelos clientes, aumentando consequentemente os seus níveis de satisfação e evitando em última análise que decidam mudar de operador.

É neste enquadramento que se definiu o principal objectivo para este trabalho: traçar o perfil dos clientes que desactivam os seus serviços (assumindo-se que um cliente que desactiva é um cliente que muda de operador, visto que actualmente um meio de comunicação móvel é considerado um bem imprescindível, sendo raros os clientes que desactivam e ficam efectivamente sem serviço). Pretende-se identificar quais são os factores que realmente têm peso nesta decisão, com vista a poder focar as acções de melhoria nas variáveis efectivamente importantes.

Sendo este o foco, procurar-se-ão identificar as variáveis relevantes para os clientes, quais as que os influenciam decisivamente na decisão de desactivar os seus serviços. Podem ter duas naturezas distintas:

- Factores comportamentais
  Estes factores não estão geralmente reflectidos nos dados disponíveis (nem sempre se traduzem em variáveis registadas nos sistemas de forma directa), pelo que podem ser mais difíceis de obter.

- Factores relacionados com o serviço
  Existe uma enorme quantidade de informação disponível (muitas variáveis em vários sistemas distintos). O processo de recolha e tratamento dos dados (escolha das variáveis, extracção dos dados, cruzamento dos dados, etc..) é complexo e moroso, requer grande conhecimento das estruturas de informação existentes.

Dada a dimensão do universo total de clientes e o facto de poderem ser divididos em grupos totalmente distintos com características muito próprias, considerou-se útil analisar isoladamente cada um desses grupos (tornando o universo de dados mais manejável). Este estudo irá incidir sobre um deles: clientes residenciais (ie particulares, excluindo-se deste estudo os clientes empresariais), pós-pagos, ou seja, que recebem uma factura mensal.

2 – Obtenção dos Dados

Conforme se referiu anteriormente, este estudo incide sobre uma parte da base total de clientes: os clientes residenciais pós-pagos.
A primeira fase, obviamente imprescindível à análise que se pretende fazer, corresponde á extracção e organização de toda a informação relevante para estes clientes. Segue-se a descrição deste processo e um resumo das variáveis que foram consideradas.
Dado o grande volume de dados correspondente a estes clientes, a tarefa de os tratar na totalidade torna-se demasiado pesada para meios informáticos "normais", pelo que se optou pela extracção de amostras que sejam representativas e a partir das quais se possam tirar conclusões para a totalidade do universo.

No que diz respeito à forma como estão organizados os dados, existem duas entidades distintas representadas nos sistemas:

MSISDN ou Serviço – corresponde na prática ao número de telefone, ou seja, ao serviço prestado a cada cliente.
Um MSISDN tem muitas características ou variáveis associadas, por exemplo uma data de activação (data a partir do qual o número passou a estar disponível para ser utilizado), estado, um determinado tarifário (que pode ser de vários tipos e pode ser alterado ao longo do tempo), serviços suplementares (voice mail, dados e fax, aditivos, etc..)

Conta – Entidade ao nível da qual é calculada e emitida a facturação. A uma conta podem corresponder um ou mais serviços (que terão os seus valores agregados numa única factura).
Tal como no caso do serviço, a conta tem muitas características ou variáveis associadas: data de activação (necessariamente igual ou anterior à data de activação do primeiro serviço da conta), estado, nome e número de contribuinte do cliente, morada de facturação, serviços suplementares (por exemplo descontos, pacotes minutos), etc…

Uma vez que o objectivo do estudo é o de caracterizar os clientes que desactivam os seus serviços, deverá ser extraída informação tanto de clientes activos como de desactivos.
A primeira decisão a tomar prende-se com o nível a que deverão ser agregados os dados: serviço ou conta. No caso das contas uni-serviço (contas às quais está associado um único serviço) esta decisão não traz grandes implicações, mas no caso das contas multi-serviço (contas com mais que um serviço associado) torna-se necessário ponderar o impacto com algum cuidado. Uma possibilidade seria considerar o serviço como base, e, para cada serviço da conta, replicar toda a informação da mesma. Em alternativa pode agregar-se toda a informação ao nível da conta, considerando variáveis como o número de serviços da conta, o tipo de serviços da conta, etc… ou seja, considerando o conjunto de serviços como um todo e não analisando cada um isoladamente.
Do ponto de vista da lógica de negócio a segunda alternativa é a que faz mais sentido: um cliente pode decidir desactivar um serviço numa conta em que tenha vários simplesmente porque não precisa daquele número em particular, não querendo isso dizer que se desvincula do seu prestador para mudar para outro, ou seja, que deixa de ser cliente. Já no caso em que desactiva a conta (e consequentemente todos os serviços que lhe estão associados) pode-se considerar que existe de facto intenção de deixar os serviços prestados pelo operador e de deixar de ser seu cliente.
Face ao exposto considera-se a conta como elemento base da análise (cada conta correspondendo a um cliente), sendo por isso toda a informação disponível extraída e organizada por conta.

2.1 - Dados para extracção de amostra

Tal como já se referiu por diversas vezes, pretende-se com este estudo identificar quais os factores que influenciam o cliente de forma negativa, levando-o à desactivação, ou seja quais as variáveis com relevância no que diz respeito ao estado da conta (activa ou desactiva). Assim sendo, deverá tomar-se como base de trabalho o universo de todas as contas, activas ou desactivas. Tal não é possível uma vez que o volume de contas desactivas é muito elevado - por incluir as contas desactivadas desde o início de actividade da empresa. Por outro lado, não faz sentido estudar desactivações ocorridas há muito tempo visto que podem corresponder a padrões de comportamento entretanto alterados ou já inexistentes.
Por este motivo, serão consideradas as contas com estado desactivo, cuja data de desactivação seja posterior a 1 de Janeiro 2007, e todas as contas activas à data da extracção dos dados (Abril 2008).

Base: Contas residenciais, pós-pagas, activas à data da extracção dos dados (Abril 08), ou desactivadas entre 01 Jan 07 e a data da extracção da amostra

Não se apresentam aqui os totais de contas (nem totais obtidos por estado) por se tratar de informação muito sensível e considerada confidencial.

2.2 - Selecção da amostra

A selecção das amostras aleatórias usadas quer para obtenção do modelo quer para sua validação serão obtidas usando o processo que se descreve em seguida.

Considerou-se que seria necessário obter uma amostra de cerca de 29.000 contas, número considerado representativo em relação ao universo total (não pode ser aqui referida a % deste número sobre o total de contas, apenas se garante que é de facto uma % considerada significativa).

A cada registo (conta, independentemente do seu estado) foi atribuído um identificador único (de 1 a n, sendo n = número total de contas).

Obtiveram-se cerca de 29.000 observações pseudo-aleatórias de uma distribuição Uniforme (0,1), usando o gerador disponível no Excel, indicando como semente o valor 65536 = 2^{16} (valor elevado que garante a distribuição aproximada pretendida).

A partir destas observações obtêm-se valores pseudo-aleatórios entre 1 e n:
-> a cada observação u $\in$ (0,1) obtida pelo processo de geração aplica-se a transformação
x = parte inteira(1 + n x u)

Seleccionando os registos correspondentes a estes valores, obtém-se a amostra sobre a qual vai incidir a análise preliminar dos dados.

NOTA: O processo devolve observações repetidas (uma vez que se considera a parte inteira de cada valor gerado, é muito provável e normal que assim aconteça), pelo que se gerou um número superior de observações – 75.000 – por forma a obter os cerca de 29.000 registos pretendidos para análise (sem manipular a amostra de NPA's obtida). Resultaram deste processo 29.403 registos, considerados como base do estudo.

Verificou-se que a proporção de contas em cada estado (activa / desactiva) é a mesma (a menos de uma décima) no universo total e na amostra extraída, o que aumenta a confiança na qualidade da amostra extraída (em termos de representatividade e aleatoriedade).

3 – Análise Exploratória dos Dados

A primeira fase, obviamente imprescindível à análise que se pretende fazer, corresponde á extracção e organização de toda a informação relevante para as contas consideradas nas amostras.

A maior parte dos dados foi extraída para a totalidade das contas do segmento em análise (cliente pós-pagos, segmento residencial), sendo depois daqui extraída a informação respeitante ás contas da amostra. Assim, se houver em fase posterior necessidade de extrair novas amostras (para validação do modelo, testes, etc..) toda esta informação estará disponível (com dados de Abril 2008).

Alguns dados (como por exemplo o número de reclamações) foram extraídos apenas para as contas da amostra (pelo seu volume, os processos de extracção não correm para a totalidade do segmento). Isto significa que se houver necessidade de usar novas amostras esta informação deverá ser extraída na altura.

3.1 - Universo das Contas em análise

Tal como se referiu anteriormente, este estudo incide sobre a base de clientes residenciais pós-pagos (ou seja, clientes não empresariais com plano de tarifas mensal, com pagamento após recepção da factura).

O universo de contas base a partir do qual vai ser efectuado o estudo foi obtido através de extracções baseadas em 2 pesquisas, uma para contas activas e outra para contas desactivas (data de desactivação entre 01-01-2007 e 31-03-2008).

A cada conta está associado um conjunto de dados que a caracteriza – tarifário, score, antiguidade, localização, geográfica, histórico de pagamentos, histórico de acções de cobranças, reclamações, etc.. Cada uma destas variáveis pode ou não ser relevante em termos do que leva os clientes a desactivarem, pelo que o primeiro passo é a recolha desta informação. Os critérios para extracção de cada uma das variáveis são descritos em seguida, sendo também apresentada uma análise univariada dos dados obtidos.
Esta análise preliminar tem por objectivo fornecer uma primeira percepção sobre os dados, distribuições aproximadas, existência ou não de correlações (que como se verá é uma informação de extrema importância no que diz respeito à formulação do modelo), importância aparente de cada uma das variáveis no que diz respeito ao perfil dos clientes que desactivam. Esta análise exploratória baseia-se no cálculo de estatísticas descritivas e de representações gráficas, e tem por objectivo fornecer uma antevisão de qual o tipo de modelo que melhor se vai ajustar aos dados disponíveis e explicar os motivos que levam os clientes à desactivação.

Os dados existentes estão estruturados numa hierarquia de conta e serviços (a uma conta podem corresponder vários serviços). As variáveis que caracterizam cada entidade (conta) podem por isso estar ao nível da conta (havendo nesse caso uma relação de 1 para 1) ou ao nível do serviço (havendo nesse caso uma relação de 1 para n). Sempre que os dados estejam ao nível do serviço haverá necessidade de proceder ao seu tratamento, de acordo com regras a avaliar e definir em cada caso. Podemos por exemplo fazer uma contagem dos serviços que em cada conta tenham determinada característica, ou fazer a média de um valor associado a cada serviço, etc.. A forma de tratar e agregar os dados será decidida caso a caso (consoante o tipo de informação de que se trata) e descrita para cada variável.

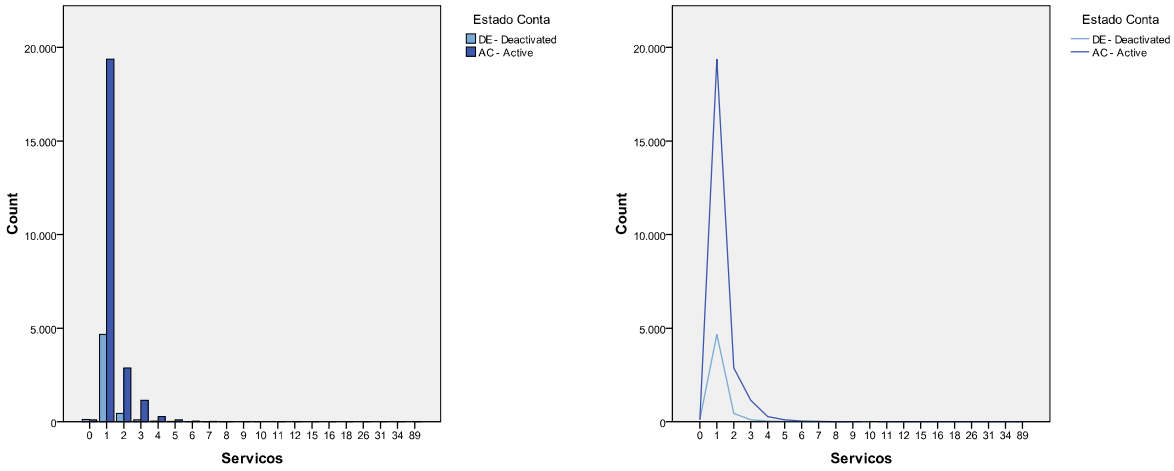
3.2 - Análise detalhada das variáveis

3.2.1 - Nº Serviços

Tal como referido anteriormente, podem estar associados a cada conta um ou mais serviços, pelo que é relevante fazer um estudo em termos do número de serviços associados a cada conta (e comparar se há diferenças relevantes entre o número de serviços associados a contas activas e contas desactivas).

NumServicos => Contagem directa do número de serviços associados à conta, independentemente do seu estado.

Os dados obtidos têm as seguintes representações gráficas:



A variável NumServicos varia entre 0 e 89, havendo poucas observações com valor superior a 8, motivo pelo qual se consideraram todas as contas com 8 ou mais serviços como pertencentes à mesma categoria (a categoria limite teve em conta o critério de se ter um mínimo de 5 observações quer para contas activas, quer para contas desactivas, o que acontece com contas até 7 serviços).
Excluíram-se as contas com 0 serviços associados pois não faz sentido incluí-las na análise (são contas com estado incoerente, possivelmente por todos os serviços terem sido desactivados, mantendo-se a conta activa por erro, ou por serem contas em fase de criação, ou por corresponderem a erros de sistema que originaram incoerência de dados).

Os dados obtidos estão sumarizados na tabela de contingência que se apresenta de seguida, onde está incluída a proporção do número de contas em cada um dos estados, para cada número de serviços associados (com 0 serviços, com 1 serviço, etc..), por forma a facilitar a análise do peso da cada categoria sobre o total:

Servicos * Estado Conta Crosstabulation

			Estado Conta		
			DE - Deactivated	AC - Active	Total
Servicos	1	Count	4672	19381	24053
		% within Servicos	19,4%	80,6%	100,0%
	2	Count	452	2881	3333
		% within Servicos	13,6%	86,4%	100,0%
	3	Count	106	1150	1256
		% within Servicos	8,4%	91,6%	100,0%
	4	Count	35	281	316
		% within Servicos	11,1%	88,9%	100,0%
	5	Count	14	111	125
		% within Servicos	11,2%	88,8%	100,0%
	6	Count	3	40	43
		% within Servicos	7,0%	93,0%	100,0%
	7	Count	3	22	25
		% within Servicos	12,0%	88,0%	100,0%
	>=8	Count	3	23	26
		% within Servicos	11,5%	88,5%	100,0%
Total		Count	5288	23889	29177
		% within Servicos	18,1%	81,9%	100,0%

O teste do Qui-quadrado, que nos permite avaliar a independência ou não das duas variáveis obteve o resultado seguinte:

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	173,136[a]	7	,000
Likelihood Ratio	194,794	7	,000
N of Valid Cases	29177		

a. 2 cells (12,5%) have expected count less than 5. The minimum expected count is 4,53.

Aos níveis de significância habituais rejeita-se H_0, ou seja, conclui-se que existe dependência entre as duas variáveis.

3.2.2 - Tarifário (Pricing Plan)

Os tarifários são uma característica do serviço, o que significa que podem estar agrupados numa mesma conta serviços com tarifários distintos. Por este motivo, começou por extrair-se o tarifário associado a cada serviço (para os serviços pertencentes às contas em análise). O número de tarifários distintos é extremamente elevado, pelo que se usou para esta análise uma classificação já existente para os mesmos: PricingPlanType, com 17 valores distintos. Cada conta foi classificada de acordo com existência ou não de algum serviço em cada uma das classes de tarifário, ou seja, para cada classe de tarifário, a conta é classificada a 0 caso não tenha nenhum serviço nessa classe, a 1 caso contrário.

A manipulação de 17 variáveis distintas torna a análise bastante complexa e dificulta a obtenção de conclusões, pelo que se optou pela utilização destas 17 variáveis para classificar as contas de acordo com o tipo de serviços associados.

Com este intuito calcularam-se o número de serviços de voz, de dados fixos e dados móveis associados a cada conta. A partir destes contadores as contas foram classificadas em tipos da forma que se segue:

Voz – Todos os serviços da conta têm tarifário GSM (Voz)
DadosFixos (ADSL) – Todos os serviços da conta são do tipo ADSL
DadosMoveis (BLM) – Todos os serviços da conta são do tipo banda larga móvel
Mista – A conta tem serviços de voz e serviços de dados

Tal como aconteceu com a variável anterior, encontraram-se dados incoerentes para algumas contas (que ou não têm serviços associados, ou cujos serviços não têm tarifário válido). Também nesta análise estas contas serão excluídas.

Os dados organizados de acordo com estas categorias resultantes dos agrupamentos efectuados são apresentados de seguida:

Estado Conta * Classificacao Crosstabulation

			Classificacao				
			DadosFixos	DadosMoveis	Mista	Voz	Total
Estado Conta	DE - Deactivated	Count	101	2558	44	2544	5247
		% within Classificacao	6,2%	20,0%	9,4%	18,6%	18,4%
	AC - Active	Count	1541	10213	423	11149	23326
		% within Classificacao	93,8%	80,0%	90,6%	81,4%	81,6%
Total		Count	1642	12771	467	13693	28573
		% within Classificacao	100,0%	100,0%	100,0%	100,0%	100,0%

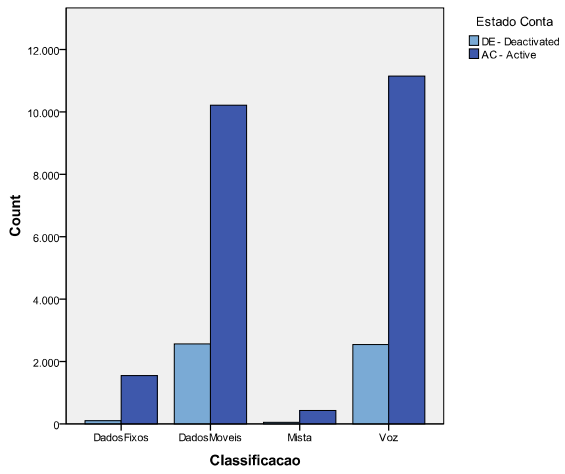
Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	212,340[a]	3	,000
Likelihood Ratio	261,476	3	,000
N of Valid Cases	28573		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 85,76.

O teste do qui-quadrado aplicado a estes dados rejeita a hipótese de independência entre as duas variáveis:

A representação gráfica dos dados é apresentada em seguida e sugere proporções distintas para contas activas versus desactivas em cada um dos tipos:

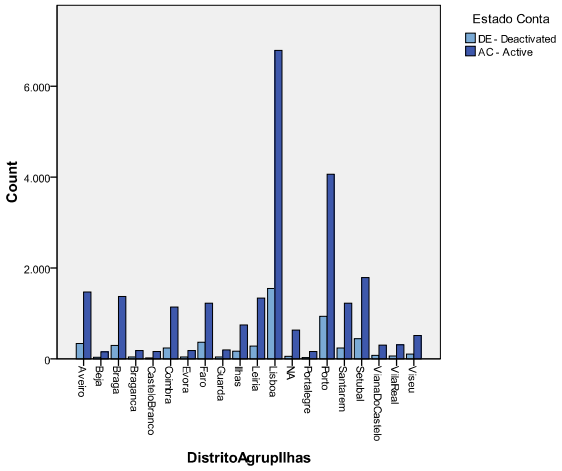


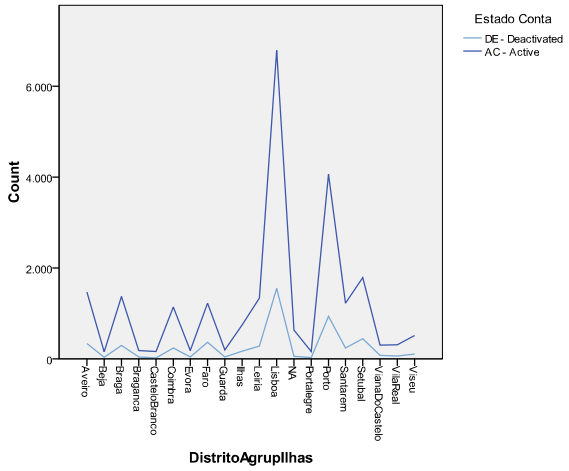
3.3.3 - Localização da conta

As contas em análise são todas do tipo pós-pagas, o que significa que recebem mensalmente uma factura, estando portanto identificadas e com uma morada associada. A informação relativa à sua localização geográfica pode revelar-se importante na definição de um padrão para os clientes mais propensos a desactivar. Por este motivo inclui-se uma análise do estado das contas da amostra versus a sua localização (distrito).

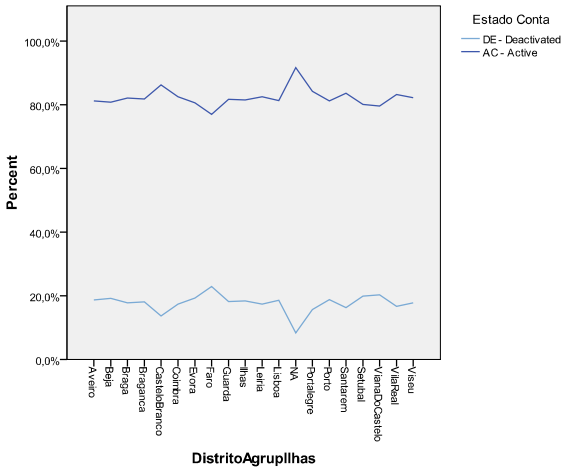
Extraiu-se todo o detalhe associado à morada, Concelho, Localidade, Distrito, Cidade. Para efeitos de agrupamento considerou-se o Distrito (obtendo-se 33 valores distintos).
Verificaram-se contagens muito baixas para alguns distritos das Ilhas dos Açores e da Madeira, pelo que se optou por agrupar todos numa única categoria: Ilhas (assume-se que o comportamento dos clientes não deve ser muito distinto de umas ilhas para as outras).

Os dados obtidos foram agrupados e são apresentados nos gráficos seguintes:





As proporções de contas activas e desactivas em cada um dos distritos são representadas no gráfico seguinte:



A maior variação verifica-se na categoria NA (contas para as quais não foi possível obter informação, correspondem a erros de registo ou erros de sincronização de dados entre sistemas). Estes registos não vão ser considerados no modelo uma vez que se tratam de dados inválidos, pelo são desde já excluídos da análise em curso e consequentemente dos quadros seguintes, onde são apresentadas as contagens por estado da conta e distrito e os resultados do teste do qui quadrado (teste à independência das variáveis:

DistritoAgrupIlhas * Estado Conta Crosstabulation

			Estado Conta		
			DE - Deactivated	AC - Active	Total
DistritoAgrupIlhas	Aveiro	Count	340	1470	1810
		% within DistritoAgrupIlhas	18,8%	81,2%	100,0%
	Beja	Count	38	160	198
		% within DistritoAgrupIlhas	19,2%	80,8%	100,0%
	Braga	Count	298	1373	1671
		% within DistritoAgrupIlhas	17,8%	82,2%	100,0%
	Braganca	Count	42	189	231
		% within DistritoAgrupIlhas	18,2%	81,8%	100,0%
	CasteloBranco	Count	26	164	190
		% within DistritoAgrupIlhas	13,7%	86,3%	100,0%
	Coimbra	Count	241	1139	1380
		% within DistritoAgrupIlhas	17,5%	82,5%	100,0%
	Evora	Count	45	188	233
		% within DistritoAgrupIlhas	19,3%	80,7%	100,0%
	Faro	Count	365	1225	1590
		% within DistritoAgrupIlhas	23,0%	77,0%	100,0%
	Guarda	Count	44	197	241
		% within DistritoAgrupIlhas	18,3%	81,7%	100,0%
	Ilhas	Count	170	750	920
		% within DistritoAgrupIlhas	18,5%	81,5%	100,0%
	Leiria	Count	283	1339	1622
		% within DistritoAgrupIlhas	17,4%	82,6%	100,0%
	Lisboa	Count	1551	6787	8338
		% within DistritoAgrupIlhas	18,6%	81,4%	100,0%
	Portalegre	Count	30	161	191
		% within DistritoAgrupIlhas	15,7%	84,3%	100,0%
	Porto	Count	941	4065	5006
		% within DistritoAgrupIlhas	18,8%	81,2%	100,0%
	Santarem	Count	240	1227	1467
		% within DistritoAgrupIlhas	16,4%	83,6%	100,0%
	Setubal	Count	445	1791	2236
		% within DistritoAgrupIlhas	19,9%	80,1%	100,0%
	VianaDoCastelo	Count	78	306	384
		% within DistritoAgrupIlhas	20,3%	79,7%	100,0%
	VilaReal	Count	63	313	376
		% within DistritoAgrupIlhas	16,8%	83,2%	100,0%
	Viseu	Count	111	513	624
		% within DistritoAgrupIlhas	17,8%	82,2%	100,0%
Total		Count	5351	23357	28708
		% within DistritoAgrupIlhas	18,6%	81,4%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	36,746[a]	18	,006
Likelihood Ratio	36,199	18	,007
N of Valid Cases	28708		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 35,41.

O teste do qui-quadrado não permite rejeitar a hipótese de independência entre estas duas variáveis, sendo portanto expectável que a localização venha a ser excluída do modelo. Opta-se pela não exclusão à partida de qualquer das variáveis disponíveis, pelo que

a variável será inicialmente considerada apesar destes resultados.

3.3.4 - Facturação

Os valores de facturação associados a cada conta são um factor muito importante para os clientes – os montantes a pagar são naturalmente muito valorizados, quer pela positiva, quer pela negativa, consoante o montante da factura esteja ou não próximo do valor que o cliente espera e considera justo face à utilização que fez dos serviços. Assim, faz sentido incluir a facturação neste estudo, com o objectivo de aferir se esta é uma variável que pesa na decisão de desactivação.

Com este objectivo foram extraídos os valores de todas as facturas associadas às contas em análise, respectiva data e valor.
Para contas activas foram consideradas as facturas entre 01-01-2007 e 31-03-2008 (em três processos de extracção por semestres, por questões de volume de dados retornados).
Para contas desactivas foi necessário obter também dados de facturação mais antigos: desde 01-01-2006 (para as contas anteriormente consideradas, ou seja, desactivadas após 01-01-2007).

Para cada conta estes valores foram agrupados da seguinte forma:

Média de toda a informação disponível
Valor do último mês
Média dos últimos 3 meses
- ⇨ para contas activas, últimos 3 meses disponíveis
- ⇨ para contas desactivas, últimos 3 meses anteriores à data de desactivação

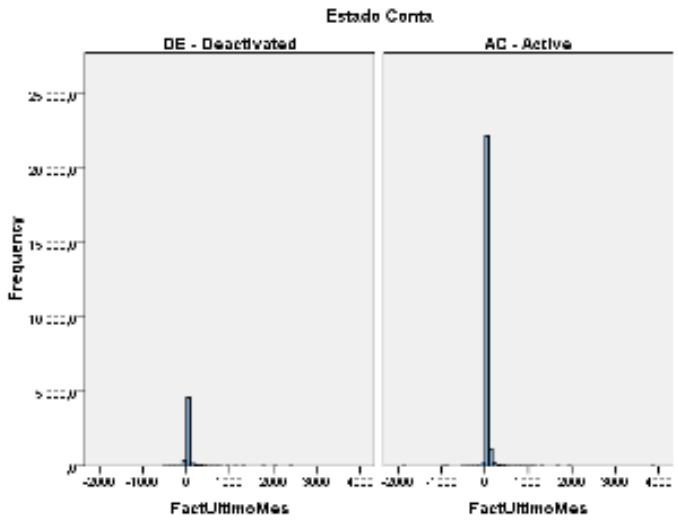
Média dos últimos 6 meses, calculados da mesma forma

O primeiro passo consiste em analisar cada uma destas variáveis separadamente e perceber se existe uma influência evidente no estado das contas.

Facturação do último mês

A experiência leva a crer que a última facturação tem um peso muito importante para o cliente: ao receber uma factura com valor muito superior ao que esperava, pode eventualmente tomar a decisão imediata de desactivar a sua conta. Por este motivo analisa-se o valor da última factura disponível.

Os gráficos e os dados seguintes mostram a existência de valores extremos muito acentuados:



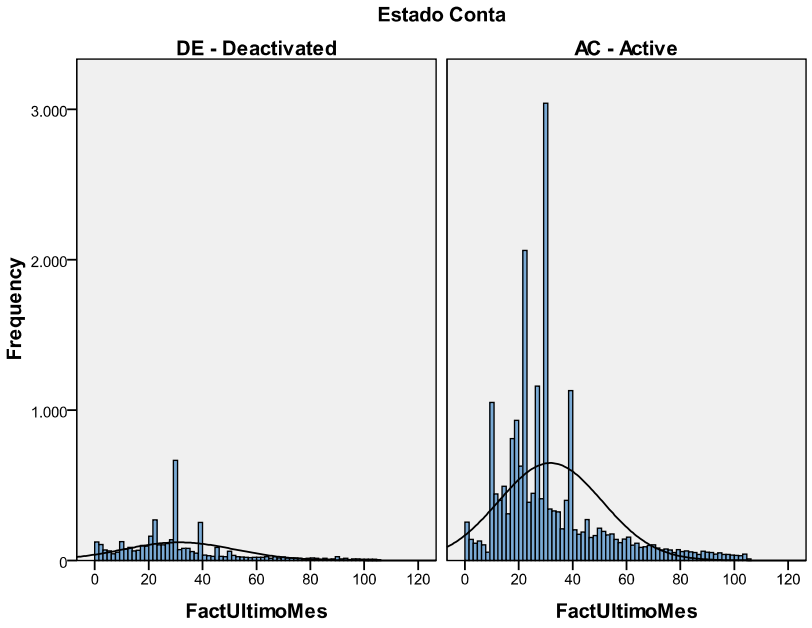
Statistics

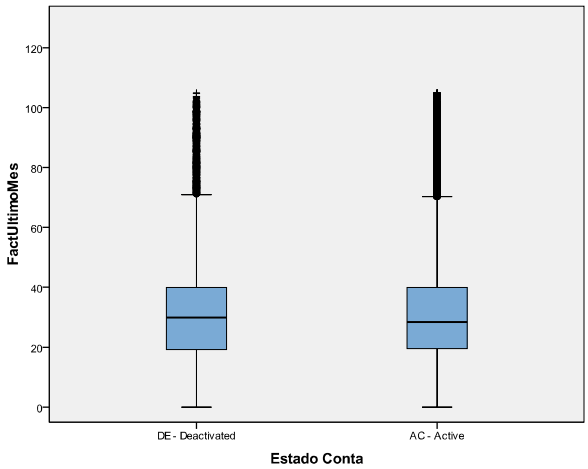
FactUltimoMes

N	Valid	29403
	Missing	0
Mean		36,19
Median		26,98
Std. Deviation		66,80
Minimum		-1865,95
Maximum		3901,71
Percentiles	1	-5,83
	5	0,00
	25	14,96
	50	26,98
	75	39,89
	95	104,96
	99	249,34

De facto, uma análise à distribuição das observações mostra que 90% dos valores se encontram entre 0€ e 104.96€, sendo o valor máximo de 3901.71€.

Para que os dados sejam mais facilmente visíveis, os mesmos vão ser representados excluindo os valores extremos de ambas as caudas. Esta simplificação é feita apenas para efeitos de representação gráfica (para permitir uma maior facilidade de percepção da distribuição), uma vez que todas as observações serão consideradas para a formulação do modelo.





Os gráficos mostram que os dois tipos de conta têm medianas muito próximas. A facturação do último mês tem uma distribuição com simetria positiva mais acentuada no caso das contas desactivas, o que indica um maior número de contas desactivas com valores baixos na última factura. A presença de muitas observações discordantes em ambos os casos mostra a existência de contas com valores de facturação muito elevados em relação às restantes observações, mas não se verificam diferenças significativas neste aspecto entre as contas activas e as contas desactivas.

Os quadros seguintes permitem comparar algumas medidas de localização e dispersão para os valores da última factura de contas activas e de contas desactivas:

Estado Conta DE - Deactivated

FactUltimoMes

N	Valid	5409
	Missing	0
Mean		37,90
Median		26,12
Std. Deviation		87,62
Minimum		-467,34
Maximum		2442,82
Percentiles	25	6,05
	50	26,12
	75	39,89

Estado Conta AC - Active

FactUltimoMes

N	Valid	23994
	Missing	0
Mean		35,80
Median		26,98
Std. Deviation		61,14
Minimum		-1865,95
Maximum		3901,71
Percentiles	25	16,77
	50	26,98
	75	39,89

A partir destes valores e das representações gráficas anteriores, podemos concluir que as distribuições não são simétricas, têm ambas forte assimetria positiva (em ambos os casos o valor da média está muito próximo do 3º quartil).
O desvio padrão é elevado para os dois conjuntos de dados, indicando uma forte dispersão dos dados em relação à média.
A média, a mediana e o 3º quartil têm valores extremamente próximos para as duas categorias de contas, há no entanto uma diferença no valor do 1º quartil – é mais baixo para contas desactivas, o que vai de encontro às conclusões já sugeridas pelos gráficos: é maior a concentração de contas com últimas facturas baixas nas contas desactivas.

Estas conclusões dizem respeito unicamente ao valor da última factura, que contrariamente ao que seria de esperar sugerem não haver normalmente uma factura de valor muito elevado imediatamente antes da desactivação. No caso das contas desactivas esta factura pode ter um decréscimo porque sendo a última, pode não dizer respeito a um período completo de facturação (depende da data de desactivação e da data em que termina o ciclo em que está inserido o cliente), ou havendo intenção de desactivar o cliente pode ter deixado de usar os seus serviços. Assim sendo, faz todo o sentido analisar a tendência de facturação dos clientes e verificar se há alterações significativas (forte subida nos valores a pagar por exemplo) que levem o cliente a querer desactivar. Analisam-se por este motivo em seguida a facturação média da conta desde a sua activação, a média dos últimos 3 e dos últimos 6 meses (no caso das contas activas, meses imediatamente anteriores à data de extracção dos dados – Abril 2008, no caso das contas desactivas, os meses imediatamente anteriores à data de desactivação). A comparação destas variáveis poderá ser útil para avaliar se diferenças no padrão de consumo levam à desactivação.

Facturação Média

O quadros seguintes apresentam as principais medidas de localização e escala e os valores extremos verificados para esta variável (média de todas as facturas associadas a cada conta):

Statistics

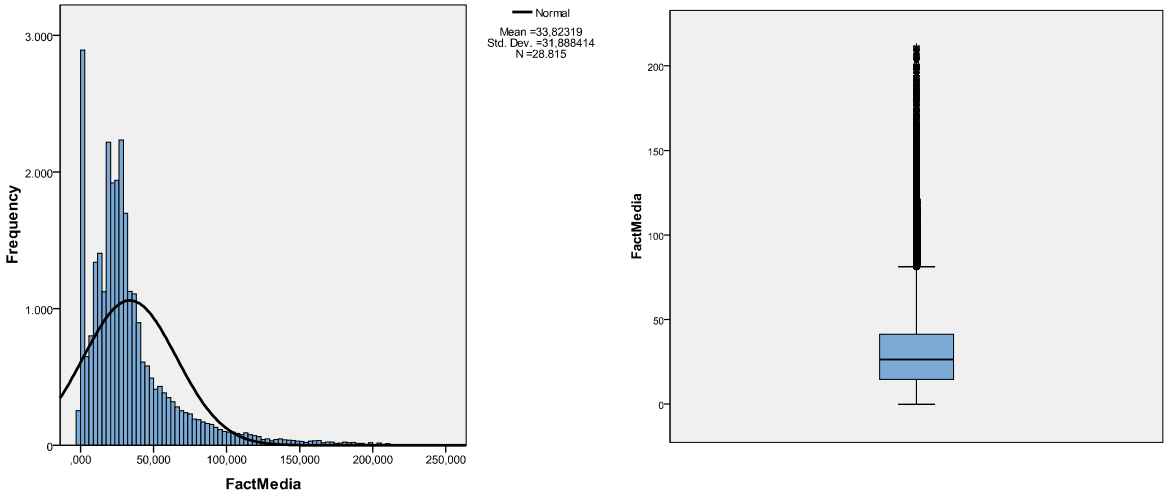
FactMedia

N	Valid	29403
	Missing	0
Mean		36,68
Median		26,31
Std. Deviation		56,08
Minimum		-300,77
Maximum		4800,49
Percentiles	1	-0,16
	5	0,00
	25	14,13
	50	26,31
	75	41,73
	95	107,85
	99	211,57

Extreme Values

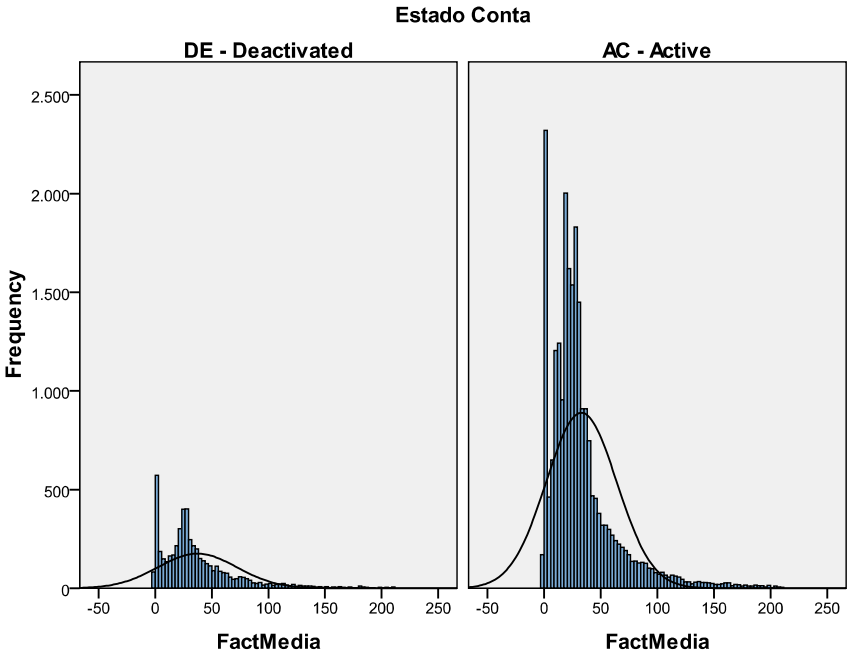
			Case Number	Value
FactMedia	Highest	1	24001	4800,49
		2	7024	1982,72
		3	18899	1419,60
		4	3159	1363,86
		5	2635	1350,62
	Lowest	1	6764	-300,77
		2	8813	-64,96
		3	9751	-62,37
		4	11713	-57,29
		5	27765	-45,44

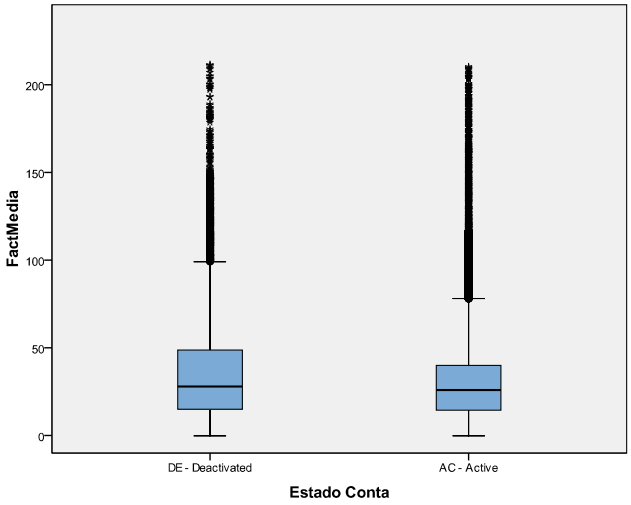
Tal como já se verificou com a variável facturação do último mês, existem valores extremos muito acentuados, o que torna a representação gráfica pouco clara. Por este motivo os gráficos seguintes representam 98% dos dados, deixando de forma as duas caudas (correspondentes a 1% das observações cada), ou seja, são representadas as observações situadas entre os valores -0.16€ e 211.57€.



Os dados apresentam forte assimetria positiva, havendo por um lado um elevado número de observações com valores abaixo da média, por outro bastantes observações discordantes à direita.

Os gráficos seguintes comparam os valores para contas activas e desactivas (mantendo a exclusão dos valores extremos):





A distribuição apresenta assimetria mais acentuada à direita no caso das contas desactivas, o que indica maior concentração de facturações elevadas para estes clientes.

Nos quadros seguintes são apresentadas as principais medidas de localização e escala para estas duas distribuições:

Estado Conta DE - Deactivated

FactMedia

N	Valid	5409
	Missing	0
Mean		41,37
Median		27,79
Std. Deviation		86,04
Minimum		-300,77
Maximum		4800,49
Percentiles	25	14,14
	50	27,79
	75	49,38

Estado Conta AC - Active

FactMedia

N	Valid	23994
	Missing	0
Mean		35,63
Median		25,95
Std. Deviation		46,68
Minimum		-62,37
Maximum		1419,60
Percentiles	25	14,12
	50	25,95
	75	40,24

O valor médio, a mediana e o 3º quartil são superiores para as contas desactivas, o que vai de encontro ao que já se observou graficamente.

Facturação Média dos últimos 3 meses

As duas variáveis anteriores permitem analisar o comportamento da conta para um período muito curto (última factura) e o comportamento ao longo de todo o seu tempo de vida (facturação média total). Poderá ter interesse a análise do comportamento médio num tempo recente face à data de extracção / data de desactivação (consoante as contas estejam activas ou desactivas), por forma a avaliar se existem padrões relevantes (e recentes) nos valores a pagar pelos clientes. Por exemplo um cliente que receba várias facturas sucessivas com valores mais altos que o habitual poderá decidir desactivar os seus serviços. Com este objectivo calcularam-se os valores médios das facturas dos últimos 3 e dos últimos 6 meses, apresentando-se em seguida a análise dos valores obtidos para estas duas variáveis.

Principais medidas de localização e escala para a facturação média dos últimos 3 meses:

Statistics

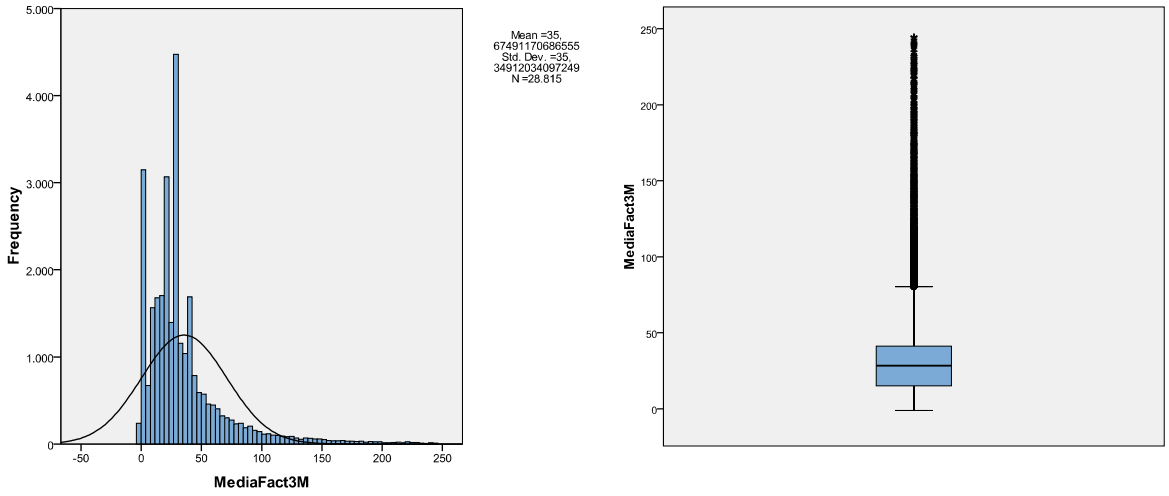
MediaFact3M

N	Valid	29403
	Missing	0
Mean		39,36
Median		28,48
Std. Deviation		72,17
Minimum		-873,61
Maximum		6400,65
Percentiles	1	-1,11
	5	0,00
	25	14,75
	50	28,48
	75	41,98
	95	118,49
	99	245,75

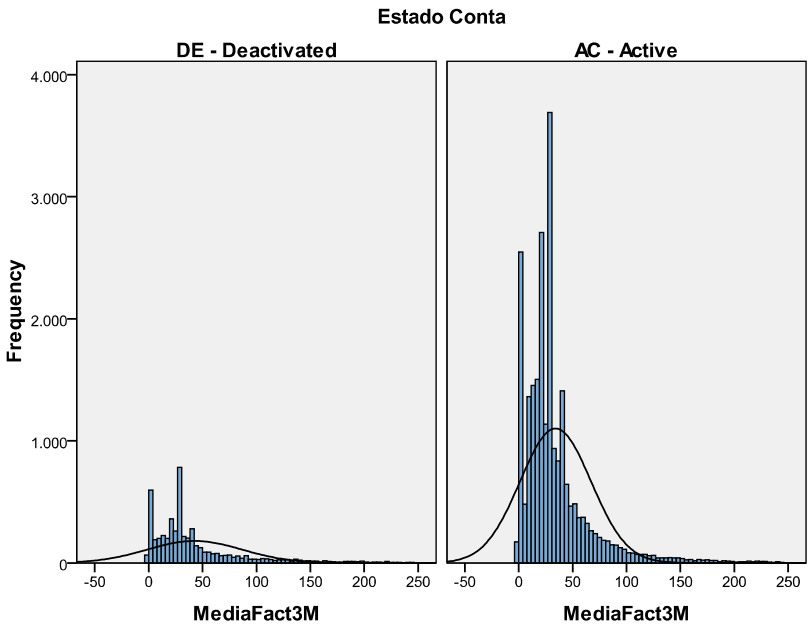
Extreme Values

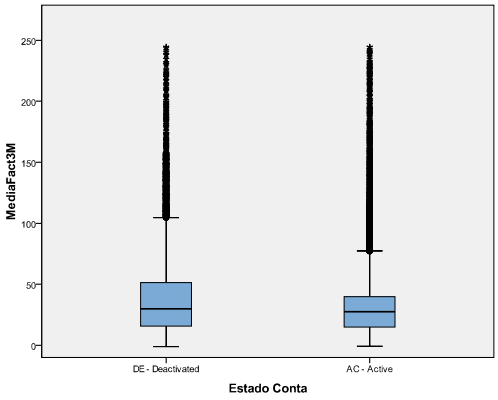
			Case Number	Value
MediaFact3M	Highest	1	24001	6400,65
		2	2635	2107,59
		3	17105	2084,37
		4	7024	2068,76
		5	17645	1709,02
	Lowest	1	14715	-873,61
		2	17188	-678,46
		3	21959	-420,36
		4	20614	-336,16
		5	17947	-317,33

São mais uma vez visíveis valores extremos (à direita) muito acentuados. Por este motivo seguir-se-á a estratégia anteriormente adoptada de representar graficamente 98% das observações, desprezando temporariamente 1% da distribuição em cada um dos seus extremos.
As representações gráficas obtidas indicam distribuições semelhantes às verificadas para as variáveis de facturação já analisadas:



Comparam-se agora as distribuições de contas activas versus contas desactivas:





Estado Conta DE - Deactivated

MediaFact3M

N	Valid	5409
	Missing	0
Mean		51,32
Median		29,90
Std. Deviation		123,35
Minimum		-873,61
Maximum		6400,65
Percentiles	25	14,49
	50	29,90
	75	53,54

Estado Conta AC - Active

MediaFact3M

N	Valid	23994
	Missing	0
Mean		36,66
Median		27,50
Std. Deviation		53,98
Minimum		-420,36
Maximum		2107,59
Percentiles	25	14,79
	50	27,50
	75	39,93

A análise destas representações gráficas e destas medidas sugere as mesmas conclusões retiradas para a facturação média: verifica-se uma assimetria positiva mais acentuada para as contas desactivas o que pode indicar que montantes de facturação altos são um factor que poderá levar os clientes à desactivação.

Facturação Média dos últimos 6 meses

Das quatro variáveis referentes a valores de facturação disponíveis, falta apenas analisar os valores da facturação média dos últimos 6 meses e confirmar se as conclusões vão de encontro às que se tiraram com as restantes variáveis.

Principais medidas de localização e escala:

Statistics

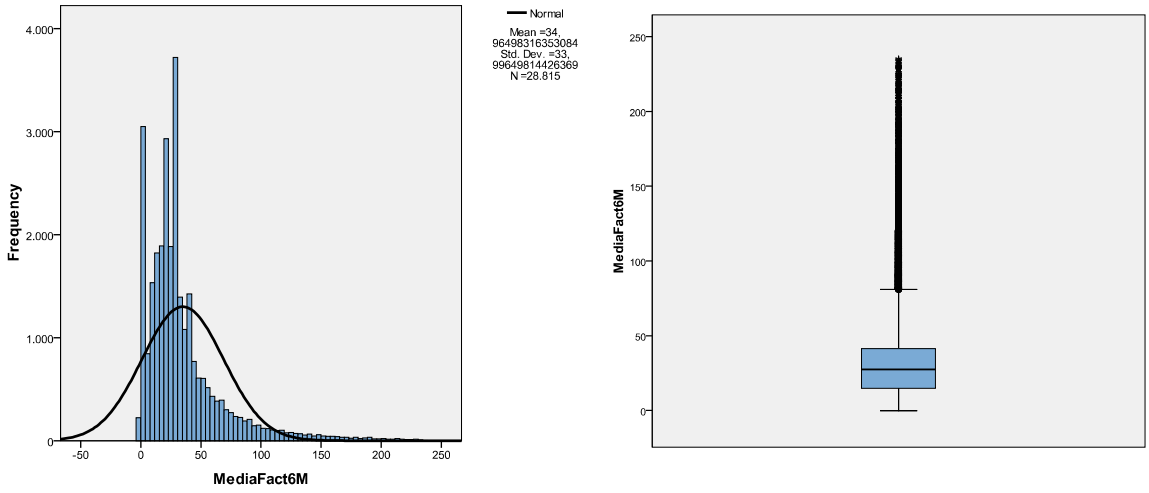
MediaFact6M

N	Valid	29403
	Missing	0
Mean		38,15
Median		27,37
Std. Deviation		60,64
Minimum		-557,60
Maximum		4800,49
Percentiles	1	-0,27
	5	0,00
	25	14,56
	50	27,37
	75	42,11
	95	114,31
	99	235,31

Extreme Values

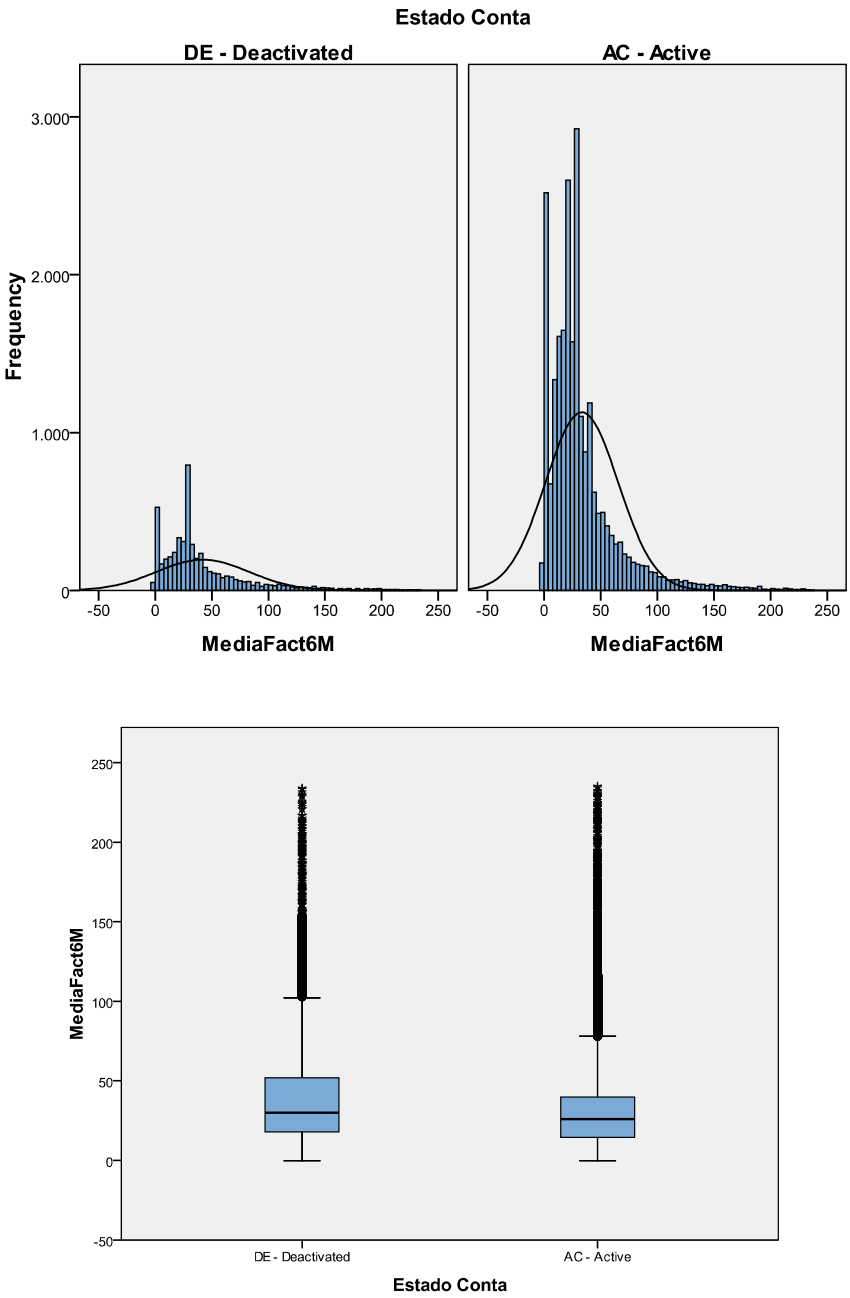
			Case Number	Value
MediaFact6M	Highest	1	24001	4800,49
		2	7024	1925,79
		3	18899	1728,81
		4	2635	1451,11
		5	23031	1357,61
	Lowest	1	18856	-557,60
		2	14901	-494,30
		3	8265	-265,76
		4	17382	-233,67
		5	15922	-207,46

Tal como foi feito anteriormente, serão consideradas para representação gráfica 98% das observações centrais da distribuição, obtendo-se os gráficos apresentados de seguida:



A distribuição destes dados é fortemente assimétrica, e são visíveis observações discordantes em elevado número à direita.

Comparam-se agora os comportamentos dos dados obtidos para contas activas e para contas desactivas:



A assimetria positiva da distribuição obtida para as contas desactivadas parece mais acentuada que para as contas activas, o que sugere que uma média de facturação elevada ao longo de 6 meses é um factor relevante que leva os clientes a quererem desactivar. Isto parece confirmar-se pela comparação das principais medidas de localização e escala de cada uma das distribuições, pois todas são bastante superiores para as contas desactivas (média, mediana, percentis e valor máximo). O desvio padrão é também bastante superior para as contas desactivas, ou seja, há aqui uma maior dispersão dos dados em torno da sua média.

Estado Conta DE - Deactivated

MediaFact6M

N	Valid	5409
	Missing	0
Mean		48,57
Median		29,90
Std. Deviation		95,57
Minimum		-233,67
Maximum		4800,49
Percentiles	25	16,69
	50	29,90
	75	53,97

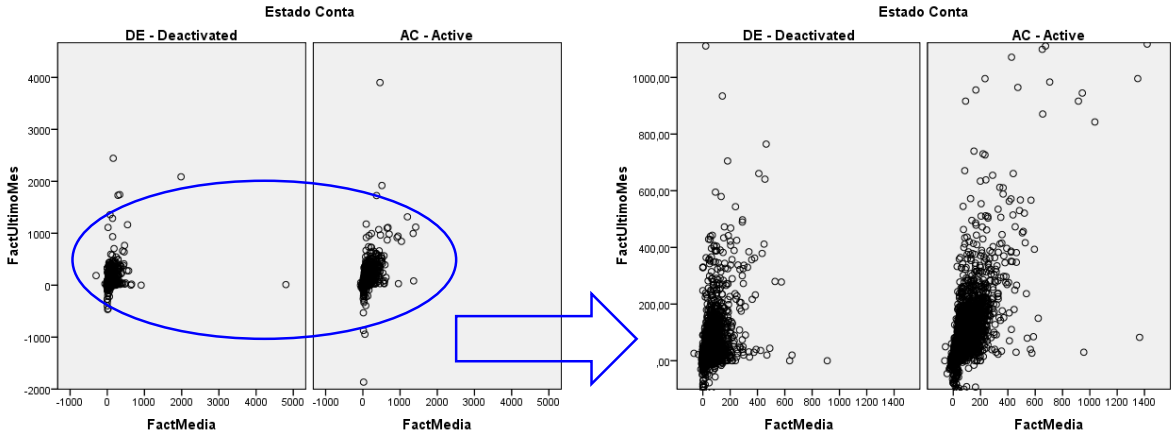
Estado Conta AC - Active

MediaFact6M

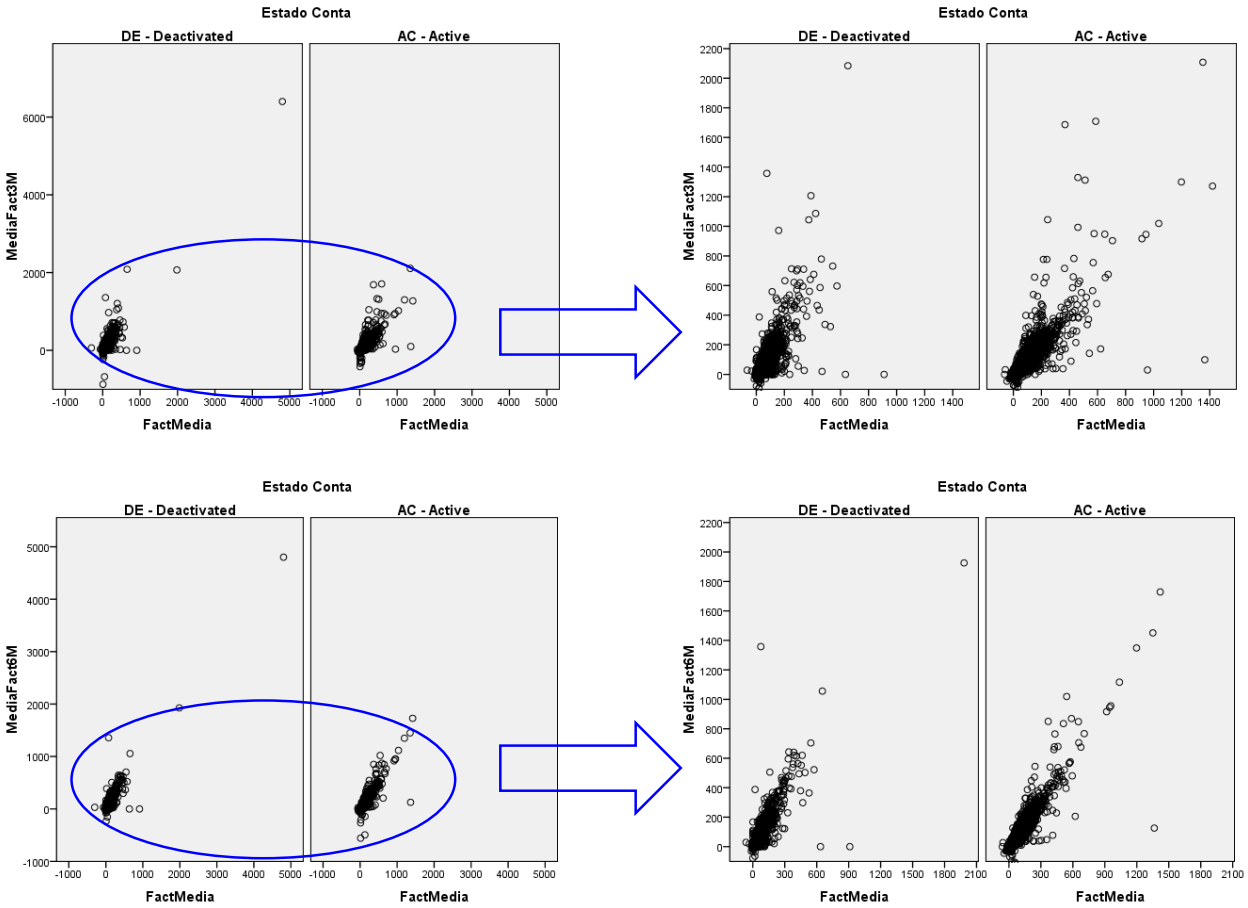
N	Valid	23994
	Missing	0
Mean		35,80
Median		25,98
Std. Deviation		49,16
Minimum		-557,60
Maximum		1728,81
Percentiles	25	14,09
	50	25,98
	75	39,95

Feita esta análise isolada de cada uma das variáveis disponíveis relativas à facturação, tem interesse proceder-se a uma análise bi-variada que permita comparar o comportamento dos valores recentes de facturação com o comportamento ao longo do tempo de vida da conta e verificar se há padrões muito distintos entre as contas activas e as contas desactivas.

Começando esta comparação pela representação gráfica conjunta da facturação média e da facturação do último mês (para contas activas e contas desactivas), não são aqui visíveis diferenças significativas:



A facturação do último mês tende a ser muito superior à facturação média, mas o padrão é semelhante para contas activas e para contas desactivas. É importante termos em conta que as datas de desactivação coincidem sempre com o fim de um ciclo de facturação, o que pode desvirtuar os dados e eventuais conclusões que se queiram tirar (o cliente pode ter deixado de utilizar os seus serviços no momento em que pede a desactivação, pelo que o valor de facturação do último mês pode não ser representativo do seu padrão recente de consumo).
Fazendo a mesma representação com os dados médios de facturação dos últimos 3 e dos últimos 6 meses obtêm-se os seguintes gráficos:



Não são visíveis padrões claramente distintos entre as contas activas e as contas desactivas.

Dada a forma como foram calculadas estas variáveis e o facto de estarem obviamente relacionadas – todas derivam dos valores de facturação da conta (valores mais antigos ou mais recentes), é possível que exista uma forte associação linear entre elas, pelo que tem interesse analisar o grau da mesma. De facto, se a associação for forte é plausível que, caso alguma destas variáveis figure no modelo, as restantes sejam excluídas.

Apresenta-se de seguida a matriz de correlações entre estas variáveis. Inclui-se também a variável que se pretende explicar – o estado da conta – apesar de, dada a sua natureza binária, não ser expectável encontrar correlações muito elevadas com as restantes (esta corresponde a uma medida de associação linear, caso as variáveis estejam fortemente associadas de forma não linear o índice de correlação será baixo, podendo por esse motivo não ser conclusivo).

A matriz indica o coeficiente de correlação para todos os pares de variáveis e respectivos graus de significância. Estão assinalados os estatisticamente relevantes:

Correlations

		Estado Conta	FactUltimo Mes	MediaFact3M	MediaFact6M	FactMedia
Estado Conta	Pearson Correlation	1	-,012*	-,079**	-,082**	-,040**
	Sig. (2-tailed)		,037	,000	,000	,000
	N	29403	29403	29403	29403	29403
FactUltimoMes	Pearson Correlation	-,012*	1	,621**	,626**	,550**
	Sig. (2-tailed)	,037		,000	,000	,000
	N	29403	29403	29403	29403	29403
MediaFact3M	Pearson Correlation	-,079**	,621**	1	,938**	,869**
	Sig. (2-tailed)	,000	,000		,000	,000
	N	29403	29403	29403	29403	29403
MediaFact6M	Pearson Correlation	-,082**	,626**	,938**	1	,930**
	Sig. (2-tailed)	,000	,000	,000		,000
	N	29403	29403	29403	29403	29403
FactMedia	Pearson Correlation	-,040**	,550**	,869**	,930**	1
	Sig. (2-tailed)	,000	,000	,000	,000	
	N	29403	29403	29403	29403	29403

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

Alguns pares de variáveis apresentam índices de correlação bastante elevados (nomeadamente a facturação média, a facturação média dos últimos 3 e dos últimos 6 meses), pelo que é expectável que a ficar alguma destas variáveis no modelo, as restantes sejam excluídas.
Tal como se esperava nenhuma das variáveis tem uma associação linear forte com o estado da conta (apesar de algumas serem consideradas estatisticamente relevantes).

3.3.5 - Acções de cobrança

As cobranças constituem como é óbvio um ponto fulcral de toda a actividade: é imprescindível controlar os pagamentos efectuados pelos clientes e as datas em que os mesmos ocorrem. Com este intuito, é associado um credit rating (CR), que define quais as acções de cobranças de que vai ser alvo e com que prazos, caso o cliente ultrapasse a data limite de pagamento de uma factura. Tipicamente começará por receber um ou mais avisos através de uma carta de cobranças, um sms ou uma chamada telefónica (outbound call). Se estes avisos não forem eficazes e a factura continuar por pagar, os serviços da conta poderão ser colocados em hotline (serviço recebe mas não faz chamadas) e finalmente, caso esta medida não produza efeito, os serviços serão desactivados temporariamente (serviço não recebe nem faz chamadas). Se ainda assim o cliente não saldar a(s) dívida(s) em aberto os serviços serão desactivados de forma definitiva e o processo passa para contencioso.
As acções e os prazos podem variar de cliente para cliente dependente da CR (Credit Rating) em que se encontrar.

O número de acções de cobranças e o tipo caracterizam cada conta do ponto de vista do cumprimento de prazos para pagamento das facturas (se as facturas de uma determinada conta forem sempre saldadas dentro do prazo, a mesma não terá cenários de cobranças associados). Assim, a análise do número de acções de cobrança permitirá concluir se é possível estabelecer uma relação entre a forma como os clientes pagam (dentro dos prazos, só depois de receber avisos de pagamento, só depois de entrarem em hotline, etc…) e a sua propensão para a desactivação.

Para este estudo só foram contabilizados os hotlines e desactivações (não estão incluídos sms, cartas ou outbound calls), pois estas são as acções que mais afectam o cliente e a utilização que faz dos seus serviços – geralmente são as mais eficazes no que diz respeito à cobrança dos valores em dívida.

Para as contas desactivas, extraíram-se os eventos desde 01-05-2006

Para cada conta os dados disponíveis foram agrupados da seguinte forma:

Número Médio de acções de cobranças (hotlines e desactivações) para todos os meses disponíveis (não faz sentido comparar totais com janelas de tempos diferentes)

- ⇨ Número de meses calculado desde a data de activação (ou 01-05-2006 se posterior), até 31-03-2008 (ou data de desactivação se anterior)

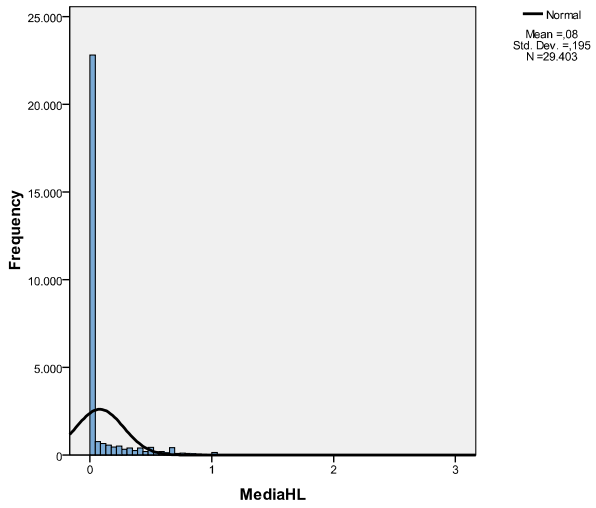
Total nos últimos 3 meses

- ⇨ para contas activas, últimos 3 meses disponíveis
- ⇨ para contas desactivas, últimos 3 meses anteriores à data de desactivação

Total nos últimos 6 meses

- ⇨ para contas activas, últimos 6 meses disponíveis
- ⇨ para contas desactivas, últimos 6 meses anteriores à data de desactivação

A primeira análise é feita às variáveis referentes ao número de hotlines por conta (primeira acção, o cliente recebe mas não faz chamadas). A média de hotlines é muito baixa (nula para a maioria das contas), conforme se pode verificar no gráfico e quadros seguintes:

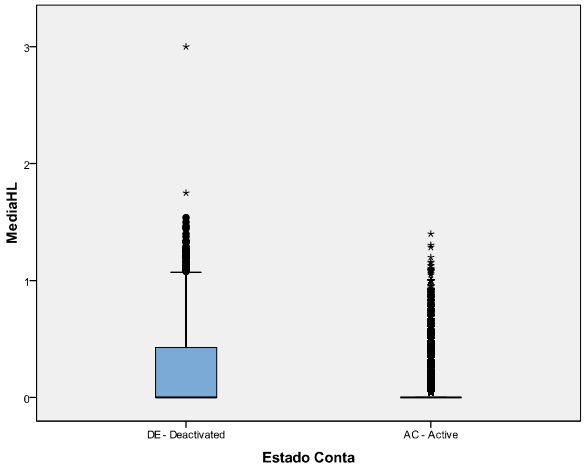


Statistics

MediaHL

N	Valid	29403
	Missing	0
Mean		0,08
Median		0,00
Std. Deviation		,20
Minimum		,00
Maximum		3,00
Percentiles	25	0,00
	50	0,00
	75	0,00

A análise das distribuições obtidas para contas activas e contas desactivas é apresentada em seguida:



As médias de hotline tendem a ser superiores para as contas desactivas, apesar de em ambos os casos a mediana ser nula (ou seja, pelo menos metade das observações tem este valor).
Para as contas activas o 3º quartil também é nulo, tendo o valor de 0.43 nas contas desactivas.

Estado Conta DE - Deactivated

MediaHL

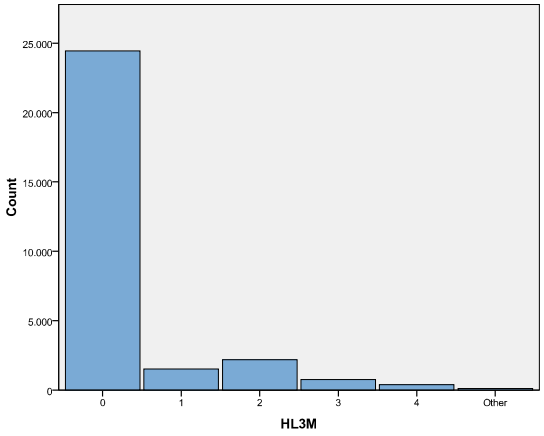
N	Valid	5409
	Missing	0
Mean		0,22
Median		0,00
Std. Deviation		,31
Minimum		,00
Maximum		3,00
Percentiles	25	0,00
	50	0,00
	75	0,43

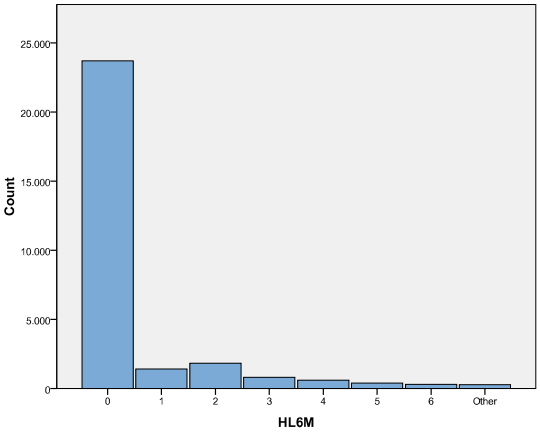
Estado Conta AC - Active

MediaHL

N	Valid	23994
	Missing	0
Mean		0,05
Median		0,00
Std. Deviation		,14
Minimum		,00
Maximum		1,40
Percentiles	25	0,00
	50	0,00
	75	0,00

A análise do número de hotlines nos últimos 3 e nos últimos 6 meses pode fornecer algum indício sobre eventuais padrões de comportamento (no que diz respeito a pagamentos) imediatamente antes da desactivação definitiva das contas.





As distribuições destas duas variáveis apresentam um comportamento muito semelhante, com a maior parte das observações - pelo menos 75% - correspondentes a zero (ou seja, 75% das contas não entrou em hotline nos últimos 6 ou 3 meses).

Statistics

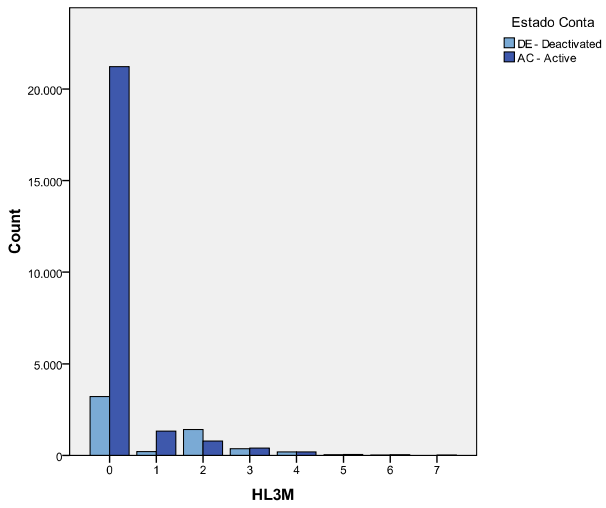
		HL3M	HL6M
N	Valid	29403	29403
	Missing	0	0
Mean		0,35	0,55
Median		0,00	0,00
Std. Deviation		,88	1,39
Minimum		,00	,00
Maximum		7,00	12,00
Percentiles	25	0,00	0,00
	50	0,00	0,00
	75	0,00	0,00

A análise seguinte corresponde às distribuições destas duas variáveis (começando pelo número total de hotlines dos últimos 3 meses) para contas activas e para contas desactivas.

Dada a natureza das variáveis (categóricas), os dados podem ser organizados em tabelas de contingência, com os resultados seguintes:

HL3M * Estado Conta Crosstabulation

			Estado Conta		
			DE - Deactivated	AC - Active	Total
HL3M	0	Count	3212	21222	24434
		% within HL3M	13,1%	86,9%	100,0%
	1	Count	204	1322	1526
		% within HL3M	13,4%	86,6%	100,0%
	2	Count	1406	781	2187
		% within HL3M	64,3%	35,7%	100,0%
	3	Count	368	391	759
		% within HL3M	48,5%	51,5%	100,0%
	4	Count	189	195	384
		% within HL3M	49,2%	50,8%	100,0%
	5	Count	23	50	73
		% within HL3M	31,5%	68,5%	100,0%
	6	Count	7	29	36
		% within HL3M	19,4%	80,6%	100,0%
	7	Count	0	4	4
		% within HL3M	,0%	100,0%	100,0%
Total		Count	5409	23994	29403
		% within HL3M	18,4%	81,6%	100,0%



As proporções de contas activas e desactivas diferem substancialmente para cada contagem de hotlines nos últimos 3 meses.

O teste seguinte rejeita a hipótese de independência entre estas duas variáveis, pelo que existe relação entre elas.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4252,772[a]	7	,000
Likelihood Ratio	3292,765	7	,000
Linear-by-Linear Association	2852,008	1	,000
N of Valid Cases	29403		

a. 2 cells (12,5%) have expected count less than 5. The minimum expected count is ,74.

A mesma análise comparativa pode ser feita para o número total de hotlines dos últimos 6 meses, com conclusões muito semelhantes:

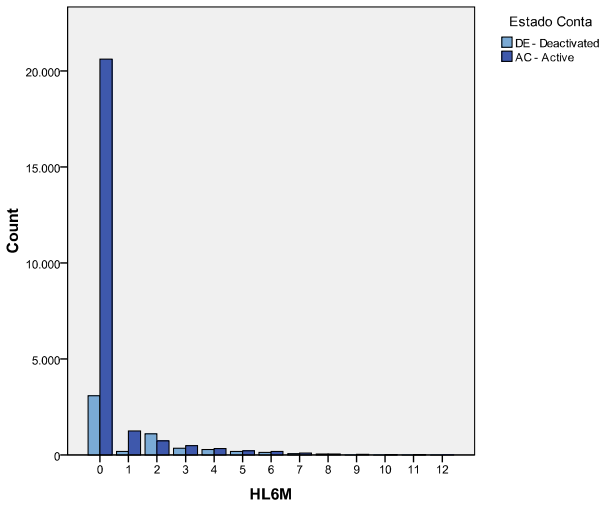
HL6M * Estado Conta Crosstabulation

			Estado Conta		
			DE - Deactivated	AC - Active	Total
HL6M	0	Count	3087	20616	23703
		% within HL6M	13,0%	87,0%	100,0%
	1	Count	176	1250	1426
		% within HL6M	12,3%	87,7%	100,0%
	2	Count	1095	735	1830
		% within HL6M	59,8%	40,2%	100,0%
	3	Count	341	478	819
		% within HL6M	41,6%	58,4%	100,0%
	4	Count	276	336	612
		% within HL6M	45,1%	54,9%	100,0%
	5	Count	186	217	403
		% within HL6M	46,2%	53,8%	100,0%
	6	Count	128	186	314
		% within HL6M	40,8%	59,2%	100,0%
	7	Count	69	91	160
		% within HL6M	43,1%	56,9%	100,0%
	8	Count	39	38	77
		% within HL6M	50,6%	49,4%	100,0%
	9	Count	4	25	29
		% within HL6M	13,8%	86,2%	100,0%
	10	Count	4	11	15
		% within HL6M	26,7%	73,3%	100,0%
	11	Count	3	8	11
		% within HL6M	27,3%	72,7%	100,0%
	12	Count	1	3	4
		% within HL6M	25,0%	75,0%	100,0%
Total		Count	5409	23994	29403
		% within HL6M	18,4%	81,6%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	3601,099[a]	12	,000
Likelihood Ratio	2881,844	12	,000
Linear-by-Linear Association	1993,393	1	,000
N of Valid Cases	29403		

a. 4 cells (15,4%) have expected count less than 5. The minimum expected count is ,74.



Dada a forma como foram obtidas e calculadas estas três variáveis, é natural que estejam fortemente correlacionadas entre si, pelo que é importante calcular os índices de correlação linear para cada par de variáveis (e de cada uma com a variável que se pretende explicar – o estado da conta):

Correlations

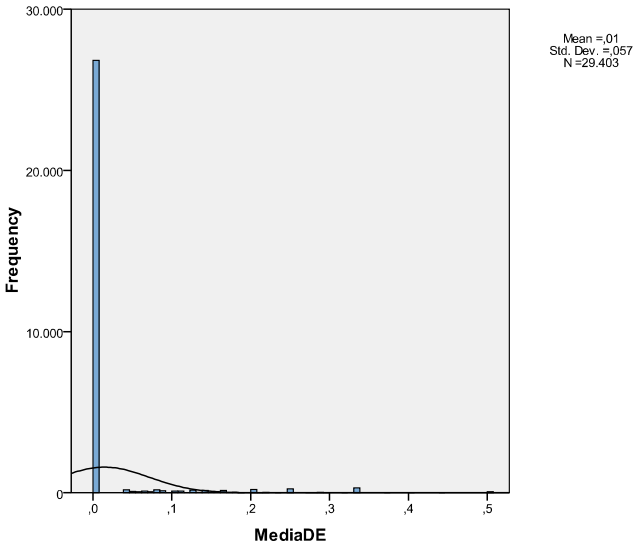
		Estado Conta	MediaHL	HL3M	HL6M
Estado Conta	Pearson Correlation	1	-,348**	-,311**	-,260**
	Sig. (2-tailed)		,000	,000	,000
	N	29403	29403	29403	29403
MediaHL	Pearson Correlation	-,348**	1	,866**	,889**
	Sig. (2-tailed)	,000		,000	,000
	N	29403	29403	29403	29403
HL3M	Pearson Correlation	-,311**	,866**	1	,918**
	Sig. (2-tailed)	,000	,000		,000
	N	29403	29403	29403	29403
HL6M	Pearson Correlation	-,260**	,889**	,918**	1
	Sig. (2-tailed)	,000	,000	,000	
	N	29403	29403	29403	29403

**. Correlation is significant at the 0.01 level (2-tailed).

É sem surpresa que se verifica que existem índices de correlação muito elevados entre as três variáveis em análise, pelo que é expectável que, a ser incluída uma delas no modelo, as restantes sejam excluídas.

Falta agora analisar o comportamento das variáveis referentes à desactivação temporária da conta (acção de cobranças mais penalizadora que o hotline na medida em que o cliente fica totalmente impedido de utilizar os seus serviços até pagar a dívida em aberto). A análise destas variáveis (média das desactivações, desactivações nos últimos 3 meses e desactivações nos últimos 6 meses) é

apresentada seguindo uma sequência semelhante à que foi apresentada para os hotlines, começando pela média de desactivações:



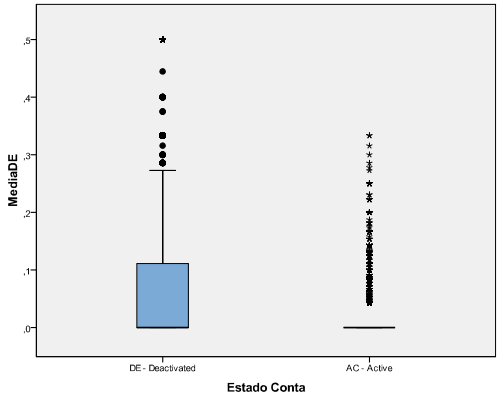
Statistics

MediaDE

N	Valid	29403
	Missing	0
Mean		0,01
Median		0,00
Std. Deviation		,06
Minimum		,00
Maximum		,50
Percentiles	25	0,00
	50	0,00
	75	0,00

Esta variável apresenta valores muito baixos, com o 3º Quartil correspondente a zero. Parece ter valores inferiores à média de hotlines, o que faz sentido, visto que muitos clientes pagam as dívidas no momento em que entram em hotline (não passando por isso pela desactivação temporária, acção posterior).

.



Estado DE - Deactivated

MediaDE

N	Valid	5409
	Missing	0
Mean		0,067
Median		0,000
Std. Deviation		,112
Minimum		,000
Maximum		,500
Percentiles	25	0,000
	50	0,000
	75	0,111

Estado Conta AC - Active

MediaDE

N	Valid	23994
	Missing	0
Mean		0,003
Median		0,000
Std. Deviation		,018
Minimum		,000
Maximum		,333
Percentiles	25	0,000
	50	0,000
	75	0,000

Comparando as distribuições entre contas activas e desactivas, conclui-se que as contas desactivas apresentam uma distribuição com maior assimetria, o que indicia valores superiores (ou seja, um maior número médio de desactivações para estas contas). De facto, as contas activas apresentam 3º quartil e médias nulos, ao passo que nas contas desactivas estas medidas apresentam um valor não nulo (mas bastante baixo).

Dado que a média não fornece conclusões claras, analisam-se em seguida o número total de desactivações dos últimos 3 e dos últimos 6 meses (no caso das contas activas dos últimos 3 e 6

meses de dados, no caso das desactivas dos últimos 3 e 6 meses imediatamente anteriores à desactivação definitiva).

Statistics

		DE3M	DE6M
N	Valid	29403	29403
	Missing	0	0
Mean		0,07	0,09
Median		0	0
Std. Deviation		,27	,31
Minimum		0	0
Maximum		2	3
Percentiles	25	0	0
	50	0	0
	75	0	0

As duas variáveis apresentam observações com valores muito baixos, com médias inferiores a 0,1 (sendo o número de desactivações dos últimos 6 meses um pouco superior ao número de desactivações dos últimos 3, como seria de esperar), e com mediana e quartis nulos.

Os dados podem ser organizados sob a forma de tabelas de contingência (que a seguir se apresentam), de acordo com o estado da conta e o número de desactivações temporárias de que foi alvo. Estas tabelas permitem verificar se as proporções são semelhantes em todas as categorias, e, através de testes de qui-quadrado cujos resultados também são apresentados, verificar se é rejeitada a hipótese de independência entre as variáveis.

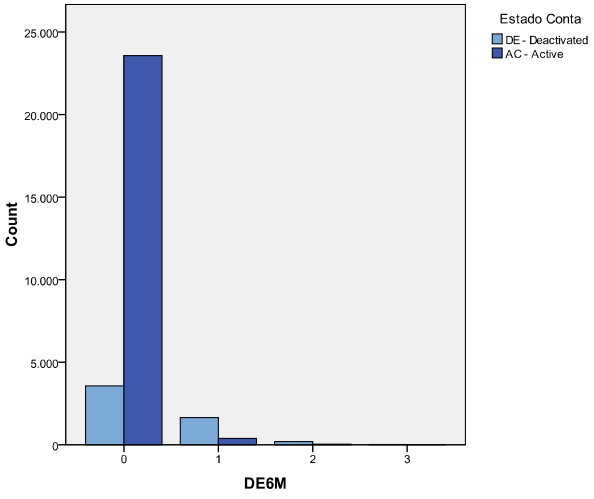
Estado Conta * DE6M Crosstabulation

			DE6M				
			0	1	2	3	Total
Estado Conta	DE - Deactivated	Count	3556	1644	192	17	5409
		% within Estado Conta	65,7%	30,4%	3,5%	,3%	100,0%
	AC - Active	Count	23567	389	36	2	23994
		% within Estado Conta	98,2%	1,6%	,2%	,0%	100,0%
Total		Count	27123	2033	228	19	29403
		% within Estado Conta	92,2%	6,9%	,8%	,1%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6511,474[a]	3	,000
Likelihood Ratio	4800,557	3	,000
Linear-by-Linear Association	5967,791	1	,000
N of Valid Cases	29403		

a. 1 cells (12,5%) have expected count less than 5. The minimum expected count is 3,50.

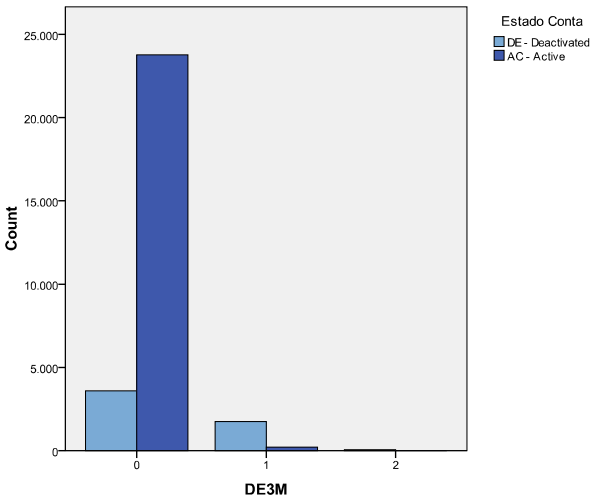


Mais de 98% das contas activas nunca foram alvo da medida de desactivação temporária, percentagem totalmente diferente da verificada para as contas desactivas, onde esta percentagem é de 65,7% (para os 6 meses anteriores à desactivação definitiva). Parece assim existir uma relação entre estas variáveis, hipótese sustentada pelo teste do qui-quadrado uma vez que este rejeita a independência entre as duas variáveis.

Semelhante conclusão se retira da análise do número de desactivações temporárias dos últimos 3 meses. A percentagem de contas que não foram alvo desta acção é substancialmente superior nas contas activas, sendo a hipótese de independência entre as variáveis rejeitada pelo teste do qui-quadrado:

Estado Conta * DE3M Crosstabulation

			DE3M			Total
			0	1	2	
Estado Conta	DE - Deactivated	Count	3596	1752	61	5409
		% within Estado Conta	66,5%	32,4%	1,1%	100,0%
	AC - Active	Count	23779	211	4	23994
		% within Estado Conta	99,1%	,9%	,0%	100,0%
Total		Count	27375	1963	65	29403
		% within Estado Conta	93,1%	6,7%	,2%	100,0%



Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	7315,887[a]	2	,000
Likelihood Ratio	5405,268	2	,000
Linear-by-Linear Association	7120,479	1	,000
N of Valid Cases	29403		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 11,96.

A proporção de contas sem desactivações temporárias é muito distinta em cada um dos estados considerados, havendo naturalmente também diferenças significativas entre estados quando analisadas as contas com 1 ou mais desactivações temporárias nos últimos 3 meses.

Tal como se fez anteriormente para as três variáveis usadas para a contabilização de hotlines, também aqui, dada a natureza e a forma como foram calculadas as variáveis analisadas para as desactivações temporárias, faz sentido avaliar o grau de correlação (linear) existente entre elas.
Incluem-se novamente os hotlines para aferir se existem associações lineares relevantes com as desactivações temporárias. Os resultados são apresentados na matriz de correlações seguinte:

Correlations

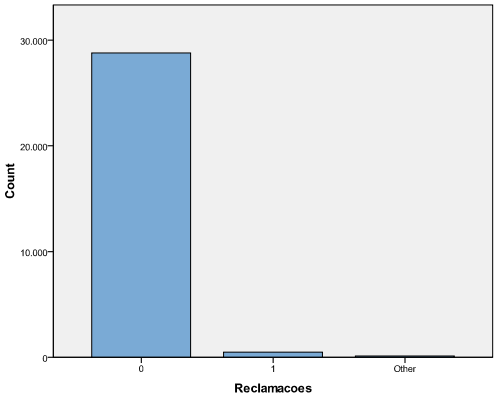
		Estado Conta	MediaHL	HL3M	HL6M	MediaDE	DE3M	DE6M
Estado Conta	Pearson Correlation	1	-,348**	-,311**	-,260**	-,442**	-,492**	-,451**
	Sig. (2-tailed)		,000	,000	,000	,000	,000	,000
	N	29403	29403	29403	29403	29403	29403	29403
MediaHL	Pearson Correlation	-,348**	1	,866**	,889**	,681**	,643**	,662**
	Sig. (2-tailed)	,000		,000	,000	,000	,000	,000
	N	29403	29403	29403	29403	29403	29403	29403
HL3M	Pearson Correlation	-,311**	,866**	1	,918**	,535**	,630**	,610**
	Sig. (2-tailed)	,000	,000		,000	,000	,000	,000
	N	29403	29403	29403	29403	29403	29403	29403
HL6M	Pearson Correlation	-,260**	,889**	,918**	1	,446**	,538**	,577**
	Sig. (2-tailed)	,000	,000	,000		,000	,000	,000
	N	29403	29403	29403	29403	29403	29403	29403
MediaDE	Pearson Correlation	-,442**	,681**	,535**	,446**	1	,836**	,831**
	Sig. (2-tailed)	,000	,000	,000	,000		,000	,000
	N	29403	29403	29403	29403	29403	29403	29403
DE3M	Pearson Correlation	-,492**	,643**	,630**	,538**	,836**	1	,917**
	Sig. (2-tailed)	,000	,000	,000	,000	,000		,000
	N	29403	29403	29403	29403	29403	29403	29403
DE6M	Pearson Correlation	-,451**	,662**	,610**	,577**	,831**	,917**	1
	Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	
	N	29403	29403	29403	29403	29403	29403	29403

**. Correlation is significant at the 0.01 level (2-tailed).

Os índices de correlação são elevados de forma generalizada. Existem pares de variáveis fortemente correlacionadas (nomeadamente as que foram calculadas com base na mesma acção de cobranças). Existem também correlações bastante elevadas entre variáveis relativas e acções diferentes (por exemplo entre a média de desactivações e a média de hotlines).
Também é de notar a associação existente entre estas variáveis e o estado da conta: é sempre assinalada como estatisticamente relevante e apresenta valores que se podem considerar bastante elevados, considerando que o estado da conta só tem dois valores possíveis.
Será por isso interessante analisar o papel que cada uma delas poderá vir a ter no modelo de forma isolada, e sobretudo de forma conjunta (sendo de esperar que não sejam todas incluídas).

3.3.6 - Reclamações

Um dos factores que se pode revelar muito importante na definição do perfil dos clientes que desactivam os seus serviços é o seu grau de descontentamento, que se pode traduzir no número de vezes que apresentaram reclamações. Cada reclamação é registada e encaminhada para uma área específica da Unidade de Operações, o GAQ – Gabinete de Apoio à Qualidade, que tem por objectivo tratar de cada caso e responder aos clientes (resolvendo sempre que possível o problema que deu origem à mesma e registando a forma como a situação foi concluída).
Os registos de reclamações podem ficar associados às contas ou aos serviços consoante faça mais sentido de uma forma ou de outra, dependendo sobretudo da natureza da mesma. Por exemplo um problema de facturação será tipicamente registado ao nível da conta (uma vez que a facturação é sempre agregada por conta), um problema com o equipamento ao nível do serviço, etc.. Por este motivo foram extraídos todas reclamações registadas desde 01/05/2006, quer ao nível do serviço quer ao nível da conta, e calculou-se com base nestes dados o número total de reclamações por conta (considerando os registos da própria conta e os registos associados a todos os serviços da conta). Os valores obtidos são representados no gráfico e sumarizados na tabela seguintes:



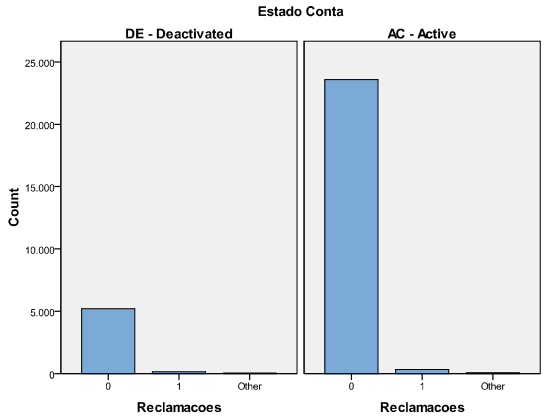
Statistics

Reclamacoes

N	Valid	29403
	Missing	0
Mean		,03
Median		,00
Std. Deviation		,212
Minimum		0
Maximum		7
Percentiles	25	,00
	50	,00
	75	,00

O número de reclamações por conta tende a ser muito baixo, com pelo menos 75% das contas sem qualquer registo.

Comparando o número de reclamações das contas activas com o das contas desactivas, obtêm-se as seguintes distribuições:



Estado Conta DE - Deactivated

Reclamacoes

N	Valid	5409
	Missing	0
Mean		,05
Median		,00
Std. Deviation		,284
Minimum		0
Maximum		5
Percentiles	25	,00
	50	,00
	75	,00

Estado Conta AC - Active

Reclamacoes

N	Valid	23994
	Missing	0
Mean		,02
Median		,00
Std. Deviation		,191
Minimum		0
Maximum		7
Percentiles	25	,00
	50	,00
	75	,00

Não são visíveis diferenças significativas nas medidas de localização e escala de cada uma das distribuições.

Dada a natureza das variáveis as mesmas podem ser organizadas numa tabela de contingência, com as contagens de contas correspondentes às categorias obtidas pelo cruzamento dos valores possíveis para as duas variáveis em estudo: estado da conta e número de reclamações. Como se obtêm muitas células com valor inferior a 5 (há poucas contas com número elevado de reclamações, em qualquer um dos estados), agregaram-se nas mesmas categorias (para contas activas e desactivas) as contas com 4 ou mais reclamações. Os dados assim obtidos são apresentados na tabela seguinte:

Estado Conta * Reclamacoes Crosstabulation

			Reclamacoes					
			0	1	2	3	>=4	Total
Estado Conta	DE - Deactivated	Count	5211	158	26	8	6	5409
		% within Reclamacoes	18,1%	32,0%	30,2%	40,0%	42,9%	18,4%
	AC - Active	Count	23579	335	60	12	8	23994
		% within Reclamacoes	81,9%	68,0%	69,8%	60,0%	57,1%	81,6%
Total		Count	28790	493	86	20	14	29403
		% within Reclamacoes	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	82,718[a]	4	,000
Likelihood Ratio	70,960	4	,000
N of Valid Cases	29403		

a. 2 cells (20,0%) have expected count less than 5. The minimum expected count is 2,58.

As proporções de contas activas e desactivas diferem bastante consoante o número de reclamações, parecendo por isso haver relação entre as duas variáveis, hipótese não rejeitada pelo teste de independência.

3.3.7 - Antiguidade

A antiguidade da conta é calculada a partir da sua data de activação e constitui uma forma de medir o vínculo do cliente à empresa, podendo por esse motivo constituir um factor importante a considerar para o perfil dos clientes mais propensos à desactivação.

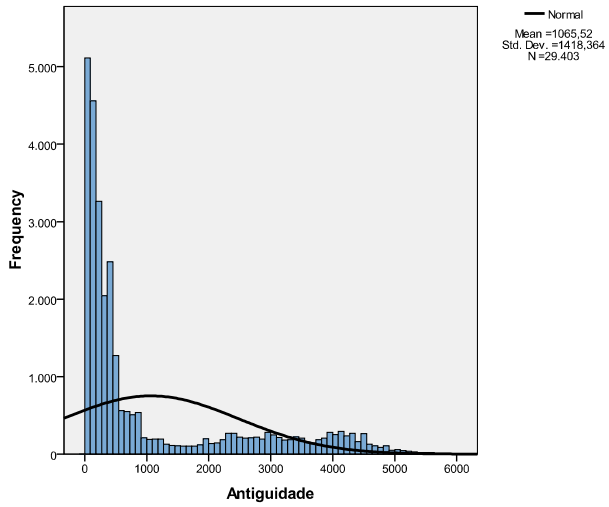
A antiguidade foi calculada da seguinte forma:

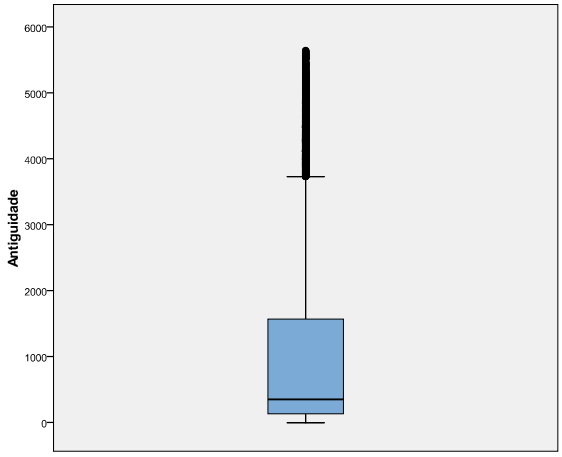
- ⇨ Contas activas
  Antiguidade = Data de extracção dos dados – Data Activação (em dias)

- ⇨ Contas desactivas
  Antiguidade = Data Desactivação – Data Activação (em dias)

Foram identificados 83 com valor de antiguidade negativa (ou seja com data de activação posterior à data de desactivação), o que se deve possivelmente a erros de registo, problemas de sincronização entre sistemas, etc… Estes registos têm assim dados incoerentes e foram por este motivo excluídos dos dados.

O gráficos seguintes representam a distribuição desta variável para todas as contas da amostra:





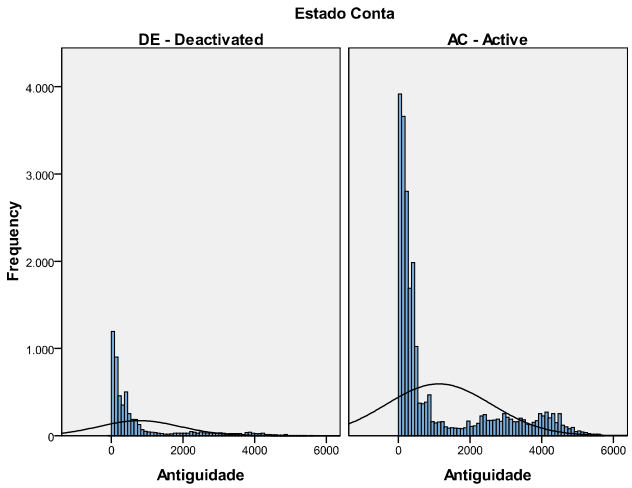
Statistics

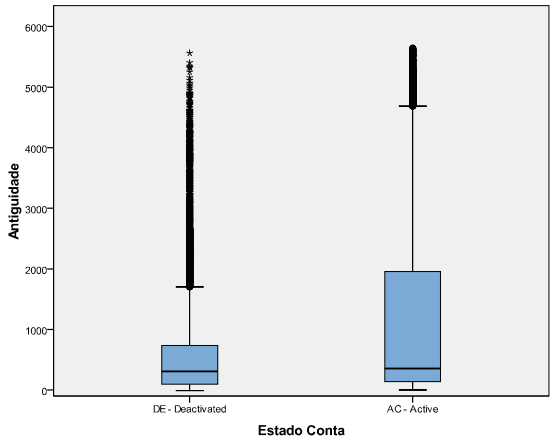
Antiguidade

N	Valid	29320
	Missing	0
Mean		1068,54
Median		352
Std. Deviation		1419,24
Minimum		1
Maximum		5644
Percentiles	25	130
	50	352
	75	1586

A distribuição apresenta forte assimetria: há um número elevado de observações com valor pequeno, mas existem também observações de valores muito elevados que influenciam a média (bastante superior à mediana).
Os dados apresentam forte dispersão em relação à média.

Comparam-se agora as distribuições da antiguidade das contas activas e das contas desactivas.





As contas activas tendem a ter antiguidade superior: a distribuição apresenta maior assimetria positiva e todas as medidas de localização e escala têm valores superiores.

Estado Conta DE - Deactivated

Antiguidade

N	Valid	5326
	Missing	0
Mean		796,51
Median		320
Std. Deviation		1144,98
Minimum		1
Maximum		5564
Percentiles	25	100
	50	320
	75	755

Estado Conta AC - Active

Antiguidade

N	Valid	23994
	Missing	0
Mean		1128,92
Median		358
Std. Deviation		1466,40
Minimum		4
Maximum		5644
Percentiles	25	138
	50	358
	75	1958,25

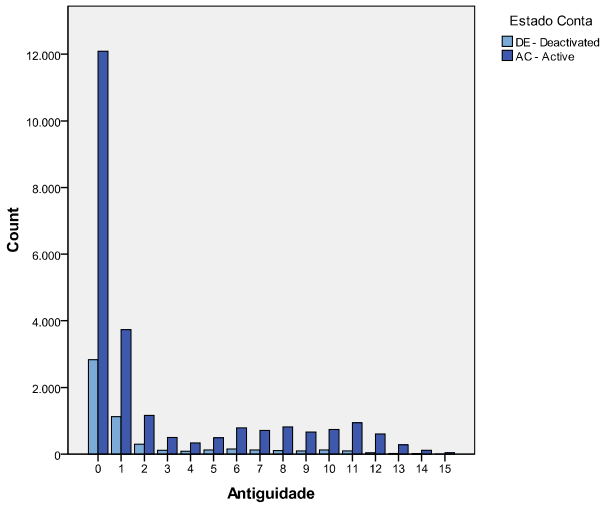
Dada a importância desta variável, optou-se por convertê-la em antiguidade em anos por forma a poder representar os dados cruzados com o estado da conta numa tabela de contingência (e testar a hipótese de independência das variáveis).

⇨ Antiguidade em anos = parte inteira (antiguidade em dias / 365)

Os dados obtidos são apresentados na tabela e gráfico seguintes:

Antiguidade * Estado Conta Crosstabulation

			Estado Conta		
			DE - Deactivated	AC - Active	Total
Antiguidade	0	Count	2830	12092	14922
		% within Antiguidade	19,0%	81,0%	100,0%
	1	Count	1122	3730	4852
		% within Antiguidade	23,1%	76,9%	100,0%
	2	Count	295	1158	1453
		% within Antiguidade	20,3%	79,7%	100,0%
	3	Count	118	499	617
		% within Antiguidade	19,1%	80,9%	100,0%
	4	Count	81	338	419
		% within Antiguidade	19,3%	80,7%	100,0%
	5	Count	120	489	609
		% within Antiguidade	19,7%	80,3%	100,0%
	6	Count	156	790	946
		% within Antiguidade	16,5%	83,5%	100,0%
	7	Count	126	713	839
		% within Antiguidade	15,0%	85,0%	100,0%
	8	Count	106	815	921
		% within Antiguidade	11,5%	88,5%	100,0%
	9	Count	92	659	751
		% within Antiguidade	12,3%	87,7%	100,0%
	10	Count	122	742	864
		% within Antiguidade	14,1%	85,9%	100,0%
	11	Count	92	937	1029
		% within Antiguidade	8,9%	91,1%	100,0%
	12	Count	38	607	645
		% within Antiguidade	5,9%	94,1%	100,0%
	13	Count	20	272	292
		% within Antiguidade	6,8%	93,2%	100,0%
	14	Count	7	118	125
		% within Antiguidade	5,6%	94,4%	100,0%
	15	Count	1	35	36
		% within Antiguidade	2,8%	97,2%	100,0%
Total		Count	5326	23994	29320
		% within Antiguidade	18,2%	81,8%	100,0%



As proporções entre contas activas e desactivas diferem consoante o número de anos da conta – nas categorias correspondentes a maior antiguidade a percentagem de contas activas aparenta ser superior.

O teste do qui-quadrado, cujo resultado é apresentado no quadro seguinte, rejeita a hipótese de independência entre as variáveis, pelo que a antiguidade parece ser relevante para o estado da conta:

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	323,339[a]	15	,000
Likelihood Ratio	366,512	15	,000
Linear-by-Linear Association	231,314	1	,000
N of Valid Cases	29320		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 6,54.

3.3.8 - Método de Pagamento

O método de pagamento associado a uma conta pode ter apenas dois valores distintos: débito directo (os montantes facturados são debitados directamente das contas dos clientes) ou outros (os clientes pagam as facturas via Multibanco, HomeBanking, pagamento nas lojas, envio de cheque, etc..). A análise desta variável pode ser relevante uma vez que os clientes que têm débito directo activo revelam normalmente um vínculo mais forte à empresa, sendo por isso interessante averiguar se têm menor propensão à desactivação.

Os dados seguintes sumarizam os dados disponíveis de acordo com as categorias das duas variáveis em estudo: o estado da conta e o método de pagamento da mesma.

Estado Conta * MetodoPagamento Crosstabulation

			MetodoPagamento		
			Direct Debit	Others	Total
Estado Conta	DE - Deactivated	Count	460	4949	5409
		% within Estado Conta	8,5%	91,5%	100,0%
	AC - Active	Count	4889	19105	23994
		% within Estado Conta	20,4%	79,6%	100,0%
Total		Count	5349	24054	29403
		% within Estado Conta	18,2%	81,8%	100,0%

A percentagem de contas com débito directo é significativamente superior nas contas activas.

A hipótese de independência entre as duas variáveis é rejeitada pelo teste do qui-quadrado, pelo que o método de pagamento poderá vir a ser importante na formulação do modelo.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	417,993[a]	1	,000		
Continuity Correction[b]	417,195	1	,000		
Likelihood Ratio	482,822	1	,000		
Fisher's Exact Test				,000	,000
N of Valid Cases	29403				

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 984,01.

b. Computed only for a 2x2 table

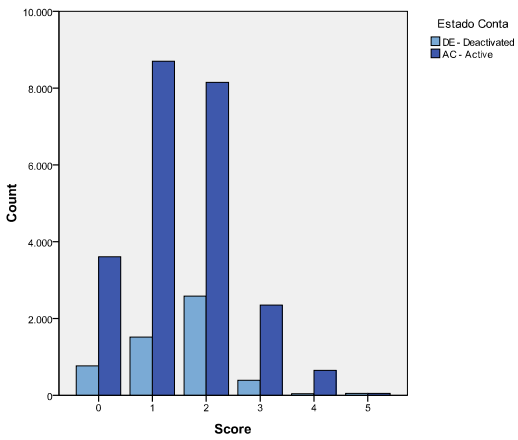
3.3.9 – Score

O score é uma variável utilizada para segmentar os clientes calculada com o valor que representam. Pode ter valores de 0 a 5 (sendo 5 o mais elevado).

Os dados organizados pelos dois factores (estado da conta versus score) distribuem-se da seguinte forma:

Score * Estado Conta Crosstabulation

			Estado Conta		
			DE - Deactivated	AC - Active	Total
Score	0	Count	764	3606	4370
		% within Score	17,5%	82,5%	100,0%
	1	Count	1516	8696	10212
		% within Score	14,8%	85,2%	100,0%
	2	Count	2584	8148	10732
		% within Score	24,1%	75,9%	100,0%
	3	Count	394	2345	2739
		% within Score	14,4%	85,6%	100,0%
	4	Count	43	651	694
		% within Score	6,2%	93,8%	100,0%
	5	Count	49	52	101
		% within Score	48,5%	51,5%	100,0%
Total		Count	5350	23498	28848
		% within Score	18,5%	81,5%	100,0%



Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	474,748[a]	5	,000
Likelihood Ratio	473,684	5	,000
Linear-by-Linear Association	18,835	1	,000
N of Valid Cases	28848		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 18,73.

O teste do Qui-Quadrado rejeita a hipótese de independência das duas variáveis, o que leva a crer que o score tenha influência no estado da conta.

3.3 – Resumo

A análise detalhada de cada variável considerada dá algumas indicações sobre a forma como se relacionam entre si e sobre a influência que poderão ter sobre o estado da conta.

A tabela seguinte constitui um resumo de todas as variáveis consideradas, do seu tipo e do seu significado. Estas serão as variáveis a considerar nas fases seguintes deste estudo:

Variável	Tipo	Descrição
Nº Serviços	Inteiro > 0	Nº Serviços associados a cada conta
Classificacao	Categorica	Voz - Todos os serviços da conta são do tipo GSM (voz) DadosFixos - Todos os serviços da conta são do tipo Dados Fixos (ADSL) DadosMoveis - Todos os serviços da conta são do tipo Dados Móveis (BLM, Banda Larga Móvel) Mista - Se a conta tem serviços de mais que um dos tipos
DistritoAgrupIlhas	Categorica	Distrito a que pertence a conta, com todos os distritos das Ilhas (Açores e Madeira) agrupados numa mesma categoria
FactUltimoMes	Contínua	Valor da última factura
FactMedia	Contínua	Valor médio de todas as facturas da conta
MediaFact3M	Contínua	Valor médio das últimas 3 facturas
MediaFact6M	Contínua	Valor médio das últimas 6 facturas
MediaHL	Contínua	Número médio de Hotlines da conta
HL3M	Inteiro ≥ 0	Número de Hotlines da conta nos últimos 3 meses
HL6M	Inteiro ≥ 0	Número de Hotlines da conta nos últimos 6 meses
MediaDE	Contínua	Número médio de desactivações temporárias da conta
HL3M	Inteiro ≥ 0	Número de desactivações temporárias da conta nos últimos 3 meses
HL6M	Inteiro ≥ 0	Número de desactivações temporárias da conta nos últimos 6 meses
Reclamacoes	Inteiro ≥ 0	Número de reclamações registadas para a conta (e respectivos serviços)
Antiguidade	Inteiro > 0	Data de Desactivação - Data de Activação (em dias), para contas desactivas Data de extracção dos dados - Data de Activação (em dias), para contas activas
MetodoPagamento	Categorica	DirectDebit se o cliente tem esta opção active, ou seja, as facturas são pagas através de débito directo Others - o cliente paga as facturas da conta através de qualquer outro método (MultiBanco, lojas, HomeBanking, etc...)
Score	Categorica	Valor do cliente (variável interna), entre 0 e 5

Conforme se referiu ao longo da análise detalhada de cada variável, foram excluídos registos com informação incoerente (por erro no registo, introdução de dados inválidos, erro na sincronização entre sistemas, etc..). Estes dados poderiam influenciar de forma incorrecta os resultados, pelo que não faz sentido serem considerados. Das exclusões sucessivas (feitas em cada variável, sempre que identificados valores errados ou incoerentes) resultaram 27929 registos que serão a base das próximas fases deste estudo.

4 – Estratégias de Modelação

Tal como se referiu por diversas vezes anteriormente, este estudo tem como principal objectivo identificar quais os factores que diferenciam as contas activas das contas desactivas, ou seja, quais as variáveis relevantes para explicar o estado de uma conta. Assim, o objectivo principal não é o estudo da incidência das desactivações, mas sim a forma como essa incidência é afectada por determinados factores. Em última análise, pretende-se obter um modelo através do qual seja possível saber qual a probabilidade de uma determinada conta (com determinadas características) vir a ser desactivada.

O estado da conta, variável que se pretende explicar (ou seja, a variável que será considerada como dependente no modelo), pode ter dois valores distintos, 0 se a conta está desactiva, 1 no caso contrário (ou seja, a conta estava activa à data da extracção dos dados), podendo as observações ser consideradas independentes umas das outras.
Esta variável pode por este motivo ser aproximada por uma variável aleatória com distribuição de Bernoulli, uma vez que esta se caracteriza por tomar apenas dois valores possíveis, associados a sucesso ou insucesso, com probabilidades p e 1 - p respectivamente.

A variável aleatória de Bernoulli é um caso particular de distribuição binomial. De facto, uma variável aleatória com distribuição binomial conta o número de sucessos em n provas de Bernoulli. Se a probabilidade de sucesso em cada prova for p, a probabilidade de se observarem k sucessos é dada por

$$p_k = \binom{n}{k} p^k (1-p)^{n-k}, k = 0,...,n$$

É evidente que para n = 1 estamos perante uma variável aleatória de Bernoulli.

As variáveis aleatórias binomiais pertencem a uma importantíssima família de distribuições, a família exponencial. Pertencem a esta família todas as distribuições (Y) cuja função densidade de probabilidade se possa escrever sob a seguinte forma:

$$f(y|\theta) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} - c(y,\phi) \right),$$

onde θ e ϕ são parâmetros escalares, a(.), b(.) e c(.) são funções reais conhecidas, sendo b(.) diferenciável e com suporte de distribuição não dependente dos parâmetros θ e ϕ.
Nesta definição θ é a forma canónica do parâmetro de distribuição e ϕ um parâmetro de dispersão.
E(Y) é dado por b'(θ), Var(Y) = a(ϕ)b''(θ)

Com efeito, conforme se referiu anteriormente, se Y ~ Bin (n, π),

$$f(y|\pi) = \binom{n}{y} \pi^y (1-\pi)^{n-y} = \exp\left(\ln\binom{n}{y} + y\ln\pi + (n-y)\ln(1-\pi) \right) = \exp\left(y\ln\left(\frac{\pi}{1-\pi}\right) + n\ln(1-\pi) + \ln\binom{n}{y} \right),$$

para y = 0, 1, ..,n.

Esta função tem assim a forma especificada pela família exponencial, com

$\theta = \ln\frac{\pi}{1-\pi}$ parâmetro canónico,

$b(\theta) = -n\ln(1-\pi) = n\ln\left(\frac{1}{1-\pi}\right)$. Ora $\theta = \ln\left(\frac{\pi}{1-\pi}\right) \Leftrightarrow \pi = \frac{e^\theta}{1+e^\theta}$, de onde se obtém facilmente que

$b(\theta) = n\ln(1+e^\theta)$.

ϕ = 1, a(ϕ) = 1

$c(y,\phi) = \ln\binom{n}{y}$

$E(Y) = b'(\theta) = n\pi$

$Var(Y) = b''(\theta) = n\pi(1 - \pi)$

Fica assim demonstrado que a distribuição pertence à família exponencial, sendo este um factor extremamente relevante no que à formulação do modelo diz respeito, conforme se verificará no capítulo seguinte.

4.1 - MLG – Modelos Lineares Generalizados

São inúmeras as situações em que existe necessidade de estudar as relações existentes entre um conjunto de variáveis, e mais especificamente, analisar a influência que uma ou mais variáveis (explicativas) possam ter sobre uma variável ou característica específica da população em estudo (variável resposta ou variável dependente).

Até há relativamente pouco tempo (meados do Século XX), este tipo de problemas era quase exclusivamente tratado através de modelos de regressão lineares, de utilização muito ampla e generalizada. Estes modelos assentam no pressuposto de a variável dependente poder ser aproximada por uma distribuição normal, o que nem sempre acontece. Tomando como exemplo o problema em análise neste estudo, em que a variável resposta tem dois valores possíveis, dificilmente se pode afirmar que a mesma pode ser bem ajustada por uma variável aleatória com distribuição Normal.
Para fazer face a situações específicas, em que a variável resposta seria bem ajustada por uma distribuição que não a normal, foram desenvolvidos vários modelos de forma isolada (o modelo complementar log-log, os modelos probit, o modelos logit, os modelos de regressão log-linear, etc…). Estes modelos têm bastante em comum: todos correspondem a uma generalização do modelo linear, permitindo relacionar uma estrutura linear com a variável resposta através de uma função (que tem como requisitos ser monótona e diferenciável). Além disso a variável resposta é sempre bem ajustada por uma distribuição pertencente à família exponencial. Têm ainda uma característica importante; assumem observações independentes (ou no mínimo não correlacionadas).
Nelder e Wedderburn introduziram e formularam em 1972 os Modelos Lineares Generalizados (MLG), metodologia que viria a agregar vários modelos estatísticos já existentes (incluindo o modelo de regressão linear) através de teoria de âmbito mais geral, ou seja, correspondente a uma classe muito ampla de modelos, dos quais são casos particulares os anteriormente referidos.

4.2 - Caracterização do Modelo

Conforme se antevê pelo exposto até ao momento, os modelos lineares generalizados pressupõem que a variável resposta seja bem aproximada por uma variável aleatória pertencente à família exponencial – condição verificada pela variável em estudo (estado da conta), como anteriormente se verificou. É também condição que as observações sejam independentes, o que acontece neste caso.
Por outro lado, estes modelos mantêm a estrutura de linearidade que caracterizam os modelos de regressão normais, mas permitem que a relação entre o valor esperado da variável resposta e o vector de covariáveis possa ser estabelecida através de uma função diferenciável, chamada função de ligação.

Para além da componente aleatória (variável resposta Y_i com distribuição pertencente à família exponencial, sendo $E(Y_i) = \mu_i$), um modelo linear generalizado é ainda caracterizado pela componente estrutural ou sistemática (preditor linear) $\eta_i = x_i^T\beta = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$, e pela relação entre este e o valor esperado μ_i, através de uma função (função de ligação) definida da seguinte forma:

$$g(E(Y_i)) = g(\mu_i) = \eta_i = x_i^T\beta$$
$$\Rightarrow \mu_i = g^{-1}(x_i^T\beta)$$

onde g - função de ligação - é monótona e diferenciável, β é um vector de parâmetros de dimensão p e x_i é uma matriz de especificação de característica p (assume-se que n = nº observações > p).

Pela sua definição não pode ser excluída à partida qualquer função de ligação na especificação do modelo a não ser que não satisfaça as condições de monotonia e diferenciabilidade já referidas. Há no entanto funções que, pelas suas propriedades e pela forma como são deduzidas, oferecem vantagens significativas quer em termos de resultados, quer em termos da interpretação do modelo, são as funções de ligação canónicas.
As vantagens obtidas com a utilização desta função são descrita adiante neste estudo, na descrição do método de estimação dos parâmetros do modelo.
A função de ligação canónica é directamente obtida da definição de função de ligação e dos parâmetros da família exponencial que caracterizam a distribuição de Y, fazendo $\theta_i = \eta_i$. Tem-se portanto $g(\mu_i) = x_i^T\beta = \eta_i$ e $\mu_i = b'(\theta_i)$, logo $\eta_i = g(\mu_i) = g(b'(\theta_i)) = \theta_i$, pelo que a função canónica g é a inversa de b'.

Neste caso, Y_i ~ Binomial (1, π_i), e conforme já vimos $b(\theta_i) = \ln(1 + e^{\theta_i})$, pelo que a inversa de b' é dada por

$$g(\theta_i) = (b')^{-1}(\theta_i) = \ln\left(\frac{\theta_i}{1-\theta_i}\right),$$

ou seja, esta é a função canónica, conhecida por função logística.
A aplicação desta função transforma a escala de probabilidades do intervalo (0, 1) para (-∞, +∞). Esta transformação da variável resposta garante que os valores preditos pelo modelo terão valores entre 0 e 1, tal como se pretende.

Isto significa que o modelo será escrito da forma

$$\ln\left(\frac{\mu_i}{1-\mu_i}\right) = x_i^T\beta = \eta_i \Leftrightarrow \mu_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} \ .$$

Estes modelos costumam designar-se por modelos de regressão logística, pela forma como se especificam e pela função de ligação utilizada.

Para que o modelo fique completamente especificado falta definir a matriz de especificação x_i, ou seja, escolher as covariáveis relevantes para a variável em estudo.

A formulação de um modelo tem por objectivo encontrar e interpretar a dependência existente entre as variáveis em estudo. Para um mesmo conjunto de dados poderão existir vários modelos que expliquem de forma satisfatória a sua estrutura, pelo que o objectivo da modelação não pode ser encontrar o modelo adequado, e sim, encontrar de entre um conjunto de modelos válidos, qual o mais adequado de acordo com critérios de qualidade, de simplicidade de interpretação, e claro, de acordo com o objectivo do estudo.
A metodologia a utilizar para a escolha das variáveis relevantes, a selecção dos modelos e estimação dos seus parâmetros (ie, estimação dos coeficientes β's associados às covariáveis) será apresentada e detalhada nos capítulos seguintes.

4.3 – Método de estimação dos parâmetros do modelo

Ajustar um modelo aos dados consiste antes de mais em estimar os parâmetros desconhecidos do modelo. Assim, depois de escolhido um determinado modelo, devem ser estimados os seus parâmetros e verificada a precisão dos mesmos. No caso dos modelos de regressão linear, são geralmente usados dois métodos distintos para encontrar estimativas dos parâmetros: o método dos mínimos quadrados e o método da máxima verosimilhança (que no caso concreto da regressão linear fornecem os mesmos estimadores).

O método dos mínimos quadrados consiste em encontrar os parâmetros β_0, β_1, …, β_p do modelo que minimizam a soma dos quadrados dos desvios entre os valores observados e os seus valores esperados, ou seja, que minimizam $S = \sum_{i=1}^{n} \left(y_i - x_i^T \beta\right)^2$.

Geralmente os estimadores de β são obtidos pela diferenciação de S em relação a cada parâmetro β_j, e resolvendo o sistema de equações simultâneas

$$\frac{\partial S}{\partial \beta_j} = 0, j = 1,...,p$$

Claro que é necessário validar se as soluções correspondem a mínimos, ou seja, se a matriz de segundas derivadas é definida positiva e identificar os mínimos globais de entre as soluções, bem como se existem mínimos locais nas fronteiras do espaço de parâmetros. No caso da regressão linear as soluções destas equações (as equações normais) correspondem sempre a mínimos.

Este método é usado de forma quase universal na estimação de parâmetros de modelos de regressão linear, porque os estimadores obtidos são centrados de variância mínima. Além disso, quando há normalidade dos dados, a soma dos quadrados dos desvios será bem aproximada por uma distribuição χ^2, o que irá permitir a utilização de testes F para calcular a significância de um modelo ou para comparar dois modelos aninhados (ie, em que um deles contém todos os termos do outro). Conforme já referido, nos modelos de regressão linear os estimadores encontrados com este método coincidem com os estimadores de máxima verosimilhança, mas tal não acontece com os modelos lineares generalizados.

As qualidades referidas para os estimadores encontrados através do método dos mínimos quadrados não se mantêm para os modelos lineares generalizados uma vez que não se mantém o pressuposto de normalidade dos dados, sendo por isso preferível usar o método da máxima verosimilhança, que consiste em encontrar os valores dos parâmetros que maximizam a probabilidade de obter o conjunto de observações disponíveis.
Assim, para utilizar este método, encontra-se a verosimilhança dos parâmetros desconhecidos do modelo, dados os valores da amostra. Esta é dada pela densidade de probabilidade da amostra observada interpretada como função dos parâmetros desconhecidos e não como função dos dados.

Conforme se viu anteriormente, se Y com função densidade de probabilidade f pertence à família das distribuições exponenciais, então f pode ser representada da seguinte forma:

$$f\left(y|\theta\right) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} - c(y,\phi)\right)$$

Supondo que a dispersão é constante, ou seja, $a(\phi) = \phi$ é conhecida, e tendo por hipótese a independência dos componentes, a log-verosimilhança é dada por

$$\ell(\beta) = \ln \prod_{i=1}^{n} f(y_i|\theta_i) = \sum_{i=1}^{n} \ln f(y_i|\theta_i) = \sum_{i=1}^{n} \left(\frac{y_i\theta_i - b(\theta_i)}{\phi} - c(y_i,\phi)\right) = \sum_{i=1}^{n} \ell_i(\beta)$$, onde θ_i é função de β

através da relação

$$g(\mu_i) = g(b'(\theta_i)) = \eta_i = x_i^T\beta$$

Esta função pode ser interpretada como o grau de verosimilhança dos parâmetros do modelo com base nos dados observados. Quanto maior for este valor, maior será a evidência a favor de um determinado conjunto de valores para os parâmetros. Esta função pode também ser vista como um sumário da informação contida nos dados sobre os parâmetros desconhecidos.

Os estimadores de máxima verosimilhança (valores que maximizam a função de verosimilhança, e consequentemente a log-verosimilhança) serão assim solução do sistema de equações de verosimilhança

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} = 0, j = 1,2,...,p$$

As soluções encontradas deverão satisfazer a condição

$\frac{\partial^2 \ell}{\partial \beta_j^2}$ é definida positiva.

Para obter as equações de verosimilhança, podemos escrever

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i}\left(\frac{\partial \mu_i}{\partial \theta_i}\right)^{-1}\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^{-1}\frac{\partial \eta_i}{\partial \beta_j}$$

Das relações anteriormente apresentadas, obtém-se facilmente que

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{\phi} = \frac{y_i - \mu_i}{\phi}$$
$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{Var(Y_i)}{\phi}$$
$$\frac{\partial \eta_i}{\partial \mu_i} = g'(\mu_i)$$
$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$$

Destas igualdades vem

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{y_i - \mu_i}{\phi}\frac{\phi}{Var(Y_i)}\frac{1}{g'(\mu_i)}x_{ij} = \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)g'(\mu_i)}$$

As equações de verosimilhança são assim dadas por

$\sum_{i=1}^{n} \frac{y_i - \mu_i}{Var(Y_i)g'(\mu_i)} x_{ij} = 0$, j = 1,2,…,p, onde $\mu_i = g^{-1}(x_i^T\beta)$, válido para qualquer valor de ϕ.

Geralmente estas equações não são lineares nos seus parâmetros, sendo por isso necessário recorrer a métodos iterativos para encontrar as suas soluções.

Quando o modelo é formulado com base na função de ligação canónica, ie, quando

$$\eta_i = g(\mu_i) = g(b'(\theta_i)) = \theta_i ,$$

obtém-se de imediato uma simplificação na expressão da log-verosimilhança de β. De facto, neste caso tem-se que $g'(\mu_i) = \frac{\phi}{Var(Y_i)}$, donde

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{Var(Y_i) g'(\mu_i)} x_{ij} = \frac{1}{\phi} \sum_{i=1}^{n} (y_i - \mu_i) x_{ij} ,$$

ou seja, as equações de verosimilhança são dadas por $\sum_{i=1}^{n} (y_i - \mu_i) x_{ij} = 0$

O facto de usar a função de ligação canónica simplifica bastante o processo e conduz a propriedades estatísticas desejáveis para o modelo, no entanto não há razão para limitar a escolha da função de ligação a esta função, uma vez que em teoria se pode utilizar qualquer uma que cumpra os requisitos de monotonia e diferenciabilidade já referidos, havendo casos em que pode ser apropriado o uso de funções de ligação não canónicas. De qualquer forma a escolha da função de ligação canónica deverá ser sempre avaliada, pois apresenta grandes vantagens.

No caso da distribuição binomial (que temos vindo a analisar), tem-se

$$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right), \quad \phi = 1$$

$$g(\mu_i) = \eta_i \Leftrightarrow \mu_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}, \eta_i = x_i^t \beta$$

Assim, as equações de verosimilhança terão a seguinte forma:

$$\sum_{i=1}^{n} \left(y_i - \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \right) x_{ij} = 0$$

Estas equações são não lineares em β, pelo que é necessário recorrer a métodos iterativos para encontrar as suas soluções.

Um dos pontos fortes dos modelos lineares generalizados é a existência de um algoritmo único para encontrar as soluções das equações de verosimilhança, baseado numa sucessão de problemas de mínimos quadrados para os quais existem técnicas numéricas bem testadas.
Este algoritmo – método iterativo de mínimos quadrados ponderados - foi sugerido por Nelder e Weldeburn (1972), aplica-se a qualquer modelo linear generalizado, havendo apenas que fazer pequenos ajustes de acordo com a distribuição de probabilidade e a função de ligação consideradas.

Em geral, qualquer método iterativo é baseado num valor inicial $\hat{\beta}^{(0)}$ e obtém uma sequência $\hat{\beta}^{(1)}, \hat{\beta}^{(2)}$, …, de acordo com um processo de actualização de $\hat{\beta}^{(t+1)}$ baseado no valor de $\hat{\beta}^{(t)}$, terminando quando o valor de $\left\| \hat{\beta}^{(t+1)} - \hat{\beta}^{(t)} \right\|$ for suficientemente pequeno, ou seja, for inferior a um valor ε pré-estabelecido.

O método iterativo usado nos MLG utiliza esta estratégia e baseia-se no método de scores de Fisher, descrito adiante.

Considerem-se as segundas derivadas de ℓ_i, tem-se

$$-E\left(\frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k}\right) = E\left(\frac{\partial \ell_i}{\partial \beta_j} \frac{\partial \ell_i}{\partial \beta_k}\right) = E\left(\frac{(y_i - \mu_i) x_{ij}}{Var(Y_i) g'(\mu_i)} \frac{(y_i - \mu_i) x_{ik}}{Var(Y_i) g'(\mu_i)}\right) =$$

$$= \frac{x_{ij} x_{ik}}{(Var(Y_i))^2 (g'(\mu_i))^2} E\left((y_i - \mu_i)^2\right) = \frac{x_{ij} x_{ik}}{Var(Y_i)(g'(\mu_i))^2}$$

que constitui o (j,k)-ésimo elemento da matriz de informação de Fisher. Em notação matricial

$$i(\beta) = \frac{X^T W X}{\phi} \text{ , com } W = \text{diag}\,(w_1, w_2, \ldots, w_n) \text{ e } w_i = \frac{\phi}{\text{Var}(Y_i)(g'(\mu_i))^2}$$

O resultado geral da normalidade assintótica dos estimadores de máxima verosimilhança fornece a aproximação

$\hat{\beta} \sim N\left(\beta, i(\beta)^{-1}\right)$, para n elevado sendo $i(\hat{\beta})$ é um estimador consistente de $i(\beta)$ quando os modelos em estudo satisfazem certas condições de regularidade (que se verificam no caso dos modelos lineares generalizados, em particular no caso da regressão logística).
Estas condições, que garantem a consistência e a normalidade assintótica do estimador $\hat{\beta}$ são descritas com o rigor adequado em Fahrmeir e Kaufmann (1985)

Seja $\ell_*(\beta) = \left(\frac{\partial \ell}{\partial \beta_1}, \frac{\partial \ell}{\partial \beta_2}, \ldots, \frac{\partial \ell}{\partial \beta_p}\right)^T$ o vector score, pretende-se resolver a equação $\ell_*(\beta) = 0$.

O método de Newton-Raphson baseia-se na seguinte regra de actualização da (t+1)-ésima iteração:

$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + (i^{(t)})^{-1} \ell_*^{(t)}$, com $\ell_*^{(t)} = \ell_*(\beta^{(t)})$ e com $i_{jk}^{(t)} = E\left(-\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k}\right)$, j, k = 1,2,…,p, ou seja, substituindo a matriz Hessiana pelo valor esperado e passando ao simétrico (no caso em que se utiliza a função canónica a matriz Hessiana corresponde ao inverso da matriz de informação de Fisher $i(\beta)$).
Esta substituição corresponde ao método de scoring de Fisher, não alterando a convergência do algoritmo e resultando uma expressão mais simples.

A expressão apresentada é equivalente a

$(i^{(t)})\hat{\beta}^{(t+1)} = (i^{(t)})\hat{\beta}^{(t)} + \ell_*^{(t)} = X^T W^{(t)} s^{(t)}$, onde s é o vector com componentes $s_i = \frac{\partial \ell_i}{\partial \eta_i}$

Chega-se assim à expressão $X^T W^{(t)} X \hat{\beta}^{(t+1)} = X^T W^{(t)} s^{(t)}$. Obtém-se assim que a estimativa de β na (t+1)-ésima iteração é dada por

$$\hat{\beta}^{(t+1)} = \left(X^T W^{(t)} X\right)^{-1} X^T W^{(t)} s^{(t)}$$

Cada passo corresponde assim à estimação de mínimos quadrados ponderados, com os valores e os pesos a mudarem a cada iteração ($W^{(t)}$ corresponde a uma matriz de pesos).
Note-se que o cálculo de $\hat{\beta}^{(t+1)}$ não depende de ϕ, pelo que o algoritmo pode-se utilizar para $\phi = 1$, (o facto de ϕ ter ou não um valor conhecido não interfere neste algoritmo).

Pelo exposto até ao momento, o algoritmo pode ser descrito em duas etapas principais:

(a) Dado $\hat{\beta}^{(t)}$ calculam-se $s^{(t)}$ e $W^{(t)}$

(b) Obtém-se $\hat{\beta}^{(t+1)}$ através da expressão $\hat{\beta}^{(t+1)} = \left(X^T W^{(t)} X\right)^{-1} X^T W^{(t)} s^{(t)}$

Para começar o algoritmo os parâmetros são geralmente iniciados com $s^{(0)} = g(Y_i)$ e $W^{(0)}$ = Matriz identidade.

É raro ocorrerem problemas de convergência, a não ser que algum dos parâmetros de $\hat{\beta}$ seja infinito, o que significa geralmente que alguma das probabilidades estimadas é igual a 0 ou a 1. Nestes casos, apesar das probabilidades ajustadas convergirem em relação ao valor esperado, os valores das

estimativas e respectivos desvios padrão tomam valores muito elevados e não podem ser considerados fidedignos. Esta situação pode ter a ver com a estrutura dos dados, com o facto de serem esparsos, ou com a função de log-verosimilhança (que pode ser muito plana ou pelo contrário pode ter uma assímptota). Assim, perante este tipo de problemas deverão ser analisadas as possíveis causas, que podem ser várias, pois nem todos os modelos têm propriedades comuns no que diz respeito a esta questão.

A maioria dos *packages estatísticos* que incluem funcionalidades para ajustamento de modelos lineares generalizados usam algoritmos baseados no método aqui descrito.
Para encontrar um modelo adequado aos dados que se pretendem analisar será utilizado o SPSS 17.0.

4.4 – Escolha das variáveis do modelo

Conforme já referido anteriormente, a definição do modelo consiste em especificar as suas três componentes: a componente aleatória, a componente sistemática e a função de ligação.
Já estão especificadas a componente aleatória e a função de ligação, falta agora especificar a componente sistemática, ou seja, escolher quais as variáveis que vão ser incluídas nesta componente. Depois de escolhidas as variáveis, os parâmetros respectivos são estimados usando o algoritmo descrito, mas este é um processo iterativo, uma vez que depois de estimados os parâmetros se pode avaliar se são ou não significativos, se trouxeram ou não melhoria da qualidade do modelo, etc…. Pode ser necessário formular, estimar e avaliar muitos modelos até se obter um considerado adequado.
Por este motivo, havendo em geral um número considerável de variáveis independentes (como é o caso neste estudo), é importante recorrer a alguma estratégia que permita identificar quais as que devem figurar no modelo, e quais, pelo contrário, devem ser excluídas.
Existem várias metodologias para obter um modelo que se ajuste bem aos dados tendo o menor número possível de parâmetros, uma vez que os modelos mais simples são normalmente mais estáveis (do ponto de vista numérico) e mais facilmente generalizáveis. Quanto maior for o número de parâmetros, maiores serão os erros (desvios padrão estimados) e maior será a dependência do modelo em relação aos valores observados.

Uma destas metodologias consiste em partir de um modelo com todas as covariáveis, e ir eliminando as que não forem significativas (controlando sempre através de testes a qualidade de ajustamento do modelo obtido em cada passo) – *backward analisys*.
Outra metodologia – *forward analisys* - consiste no processo oposto: parte-se do modelo mais simples (modelo nulo) e testa-se a inclusão em cada passo das variáveis ainda não incluídas (verificando a qualidade do modelo resultante depois da inclusão de cada uma delas separadamente), escolhendo-se a variável mais significativa. Repete-se o processo até que as variáveis ainda não incluídas no modelo tenham nível de significância inferior ao pretendido (ou já não haja variáveis fora).
Há ainda um terceiro método – *stepwise analisys* – que consiste no anterior, verificando em cada passo a importância de cada covariável já presente no modelo, uma vez que a entrada de uma determinada covariável pode ter como efeito tornar outras (incluídas em passos anteriores) dispensáveis, havendo por isso necessidade de avaliar a sua remoção.

A "importância" de uma variável é definida com base numa medida da significância estatística do coeficiente associado à mesma. A estatística usada depende dos pressupostos do modelo. Por exemplo, na regressão linear são utilizados testes F uma vez que os erros têm distribuição aproximadamente normal. Na regressão logística os erros têm distribuição aproximadamente binomial, pelo que a significância é aferida através de teste qui-quadrado sobre a razão das log-verosimilhanças. Assim, em cada passo do processo, a variável mais importante do ponto de vista estatístico é a que produz maior alteração na log-verosimilhança relativa a um modelo que não contenha a variável, ou seja, aquela para a qual resultar um maior valor para a estatística baseada na razão das log-verosimilhanças.

Independentemente do método escolhido, deverá ser feita uma avaliação da qualidade do modelo resultante. Esta avaliação deverá sempre ter em conta várias vertentes: qualidade do ajustamento aos dados, objectivo do estudo, relevância e grau de precisão dos parâmetros, número de parâmetros, etc…

Existem várias metodologias que permitem aferir se um modelo pode ou não ser considerado adequado, de acordo com os objectivos do estudo e com critérios de precisão pré-estabelecidos. Estas metodologias assentam essencialmente em testes de qualidade do modelo (cálculo de medidas com propriedades conhecidas que permitem comparar modelos sob determinadas condições) e em testes sobre cada um dos seus parâmetros (testes de hipóteses e obtenção de intervalos de confiança para aferir se cada parâmetro é relevante para o modelo e se tem o grau de precisão desejado) e da análise dos resíduos (que permite avaliar a discrepância entre os valores observados e os valores esperados, calculados através da aplicação do modelo).
O conjunto de conclusões resultantes destes vários processos permitirá identificar de entre vários modelos possíveis qual o mais adequado aos dados disponíveis e ao problema em estudo. Realça-se aqui o que já foi referido anteriormente: para um mesmo conjunto de dados poderão ser sempre identificados vários bons modelos, pelo que o objectivo não pode ser identificar o modelo correcto, e sim, escolher de entre vários possíveis, qual o mais adequado. Por vezes a escolha poderá não ser clara, uma vez que se é certo que a adição de um novo termo ao modelo irá aumentar o seu ajustamento, este aumento poderá não ser significativo de forma a compensar o aumento de complexidade do modelo (implícito em cada novo termo). Assim, é provável que se identifiquem (através dos métodos atrás referidos) um conjunto de modelos semelhantes no que diz respeito às suas qualidades estatísticas, não havendo por isso à partida um critério único de escolha do mais adequado – em princípio, se não se puder fazer uma distinção clara em termos da qualidade estatística (ie, não se puder afirmar de forma clara e inequívoca que um é mais preciso que outro), a escolha recairá sempre sobre o mais simples, ou seja, o que tiver menos termos, uma vez que a simplicidade tem vantagens claras ao nível da estabilidade numérica do modelo, da interpretação dos resultados, da diminuição dos erros associados ao registo de observações de variáveis e de facilitar a recolha de dados caso se pretenda alimentar o modelo para previsões futuras.

4.5 – Qualidade do modelo

4.5.1 – Inferência sobre os parâmetros do modelo

Tal como já se referiu, um ponto essencial na avaliação do modelo reside na precisão das estimativas para os seus parâmetros: depois de encontradas as estimativas para os parâmetros do modelo, é imprescindível avaliar a sua qualidade, ou seja, o seu grau de correcção.
Esta avaliação pode ser feita com base em intervalos de confiança para cada um dos parâmetros, uma vez que a amplitude destes intervalos fornece uma medida da exactidão com que se pode fazer inferência. Também é importante ter a possibilidade de efectuar testes de hipóteses sobre cada um dos parâmetros, para avaliar se são significativos e se devem ou não manter-se no modelo. Para que tal seja possível, é necessário recorrer à distribuição de amostragem de $\hat{\beta}$.
Não é em geral possível obter as distribuições exactas para os estimadores de máxima verosimilhança dos β's, havendo por isso necessidade de recorrer a resultados conhecidos da teoria assintótica, que se verificam quando os modelos em estudo satisfazem certas condições de regularidade, condições essas que são satisfeitas nos modelos lineares generalizados e em particular no caso da regressão logística.

Conforme se referiu anteriormente, para n suficientemente grande, a distribuição aproximada dos estimadores de máxima verosimilhança $\hat{\beta}$ é

$$\hat{\beta} \sim N\left(\beta, i(\hat{\beta})^{-1}\right) \text{ com } i(\hat{\beta}) = \frac{X^T \hat{W} X}{\phi}$$, e $\hat{W}$ calculado no EMV $\hat{\beta}$.

A partir deste resultado podem obter-se a variância e a co-variância de $\hat{\beta}$.

Os elementos da diagonal da matriz $\left(X^T \hat{W} X\right)^{-1} \hat{\phi}^{-1}$ correspondem às variâncias estimadas, sedo por outro lado as co-variâncias fornecidas pelos termos fora da diagonal.

Estes resultados permitem de forma simples testar a hipótese de nulidade de um parâmetro β_j (H_0: β_j = 0), ou seja, testar se a variável respectiva deve ou não ser incluída no modelo.

Por outro lado, para a componente β_j de β obtém-se o intervalo de confiança com grau 1 - α

$$\hat{\beta} \pm z_{1-\alpha/2}\sqrt{\left((X^T\hat{W}X)^{-1}\right)_{j,j}}$$

Podem ocorrer problemas numéricos no ajustamento de um modelo de regressão logística, causados por estruturas existentes nos dados nem sempre possíveis de detectar na análise inicial univariada. O problema pode não estar relacionado com a correcção do modelo ou com a forma como foi especificado, mas sim com o efeito que alguns padrões nos dados podem ter no cálculo dos estimadores para os parâmetros (cálculo esse que conforme já se viu é normalmente baseado em métodos numéricos, e não em métodos exactos). Nem sempre as estimativas dos coeficientes revelam estes problemas, mas geralmente os seus desvios padrão (estimados) – que nestes casos tomam valores muito elevados - apontam para estruturas de dados "problemáticas", que podem ter várias causas (categorias nulas quando organizados em tabelas de contingência, variáveis correlacionadas, dados esparsos, etc..).
Assim, o desvio padrão estimado constitui normalmente o melhor indício de problemas numéricos que impedem a convergência do método iterativo já descrito.
Dependendo do tipo de problema identificado, deverão ser tomadas medidas (que poderão passar pela reorganização dos dados, por exemplo segmentação, agrupamentos, eliminação de variáveis correlacionadas, etc…) com o objectivo de garantir convergência do algoritmo (e portanto estimadores válidos para os parâmetros do modelo).
Este problema (e descrição das suas possíveis causas) é mais detalhado em ponto específico no fim do capítulo.

4.5.2 – Função Desvio

Um dos critérios de avaliação do modelo mais amplamente utilizado assenta no valor de uma medida usualmente denominada de *deviance* (ou desvio), baseada na função de verosimilhança do modelo em estudo. Conforme se viu anteriormente, a função de verosimilhança sumariza a informação que os dados fornecem acerca dos parâmetros desconhecidos do modelo. O valor da verosimilhança calculada nos estimadores de máxima verosimilhança pode por este motivo ser usado para aferir até que ponto os dados são bem ajustados pelo modelo em estudo (modelo corrente).

Esta estatística não pode ser usada de forma directa uma vez que não é independente do número de observações disponíveis, pelo que será sempre necessário comparar o modelo corrente (que se pretende avaliar) com um modelo que se possa tomar como base de referência – geralmente o modelo saturado (com número de parâmetros igual ao número de observações). Desta forma o desvio pode ser considerado como uma medida de comparação entre modelos uma vez que pode ser visto como medida de discrepância entre o modelo saturado e qualquer outro modelo intermédio, sendo calculado como razão da log-verosimilhança dos dois modelos que se pretendem comparar (o modelo corrente e o modelo saturado).

A log-verosimilhança do modelo saturado (calculada nos estimadores de máxima verosimilhança obtidos) será maior que qualquer função de verosimilhança destas observações (com a distribuição assumida e a mesma função de ligação), uma vez que fornece a descrição mais completa dos dados.
A medida utilizada baseia-se na estatística de Wilks, e pode ser formulada da seguinte forma:

Sejam M_K e M_P os dois modelos que se pretendem comparar, com k e p parâmetros respectivamente ($p \leq k$), $M_K \subset M_P$, ou seja, todos os termos de M_P estão em M_K.

$$\eta_{i,MK} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + ... + \beta_k x_k$$
$$\eta_{i,MP} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

Considere-se a partição de β em (β_{MP}, $\beta_{M\ K\backslash P}$), com $\beta_{MP} = \beta_0, \ldots, \beta_P$ e $\beta_{M\ K\backslash P} = \beta_{P+1}, \ldots, \beta_K$.

Pretende-se testar a hipótese $H_{0:}\ \beta_{M\ K\backslash P} = 0$ versus $H_{1:}\ \beta_{M\ K\backslash P} \neq 0$.

A estatística que permite testar estas hipóteses é dada por

$$W = -2\ln\frac{\max_{H_0} L(\beta)}{\max_{H_0 \cup H_1} L(\beta)} = 2\left(\ell(\hat{\beta}) - 2\ell(\hat{\beta}_{MP})\right)$$

Esta estatística tem no caso normal, sob H_0, distribuição de um χ^2_{k-p}, e sob algumas condições de regularidade, é aproximada por esta distribuição.

Este teste pode ser feito para um conjunto de parâmetros (tal como aqui formulado), ou para testar a significância de um único parâmetro (ou seja, para testar se esse parâmetro deve ou não ser incluído no modelo).

Conforme já vimos, no caso dos modelos lineares generalizados, a log-verosimilhança é dada por

$$\ell(\beta) = \sum_{i=1}^{n} \ell_i(\beta), \text{ com } \ell_i(\beta) = \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)$$

No caso dos modelos encaixados, a estatística W atrás formulada tem, sob H_0, distribuição assintótica de um χ^2_{k-p}, o que permite basear a inferência (e o processo de avaliação dos modelos em análise) sobre o modelo saturado.

As funções de log-verosimilhança para o modelo saturado e para o modelo em análise (modelo corrente) podem ser avaliadas nos respectivos estimadores de máxima verosimilhança (dependentes de $\tilde{\theta}$ e de $\hat{\theta}$). Se o modelo corrente se adaptar bem aos dados, $\ell(\tilde{\theta})$ deverá ser equivalente a $\ell(\hat{\theta})$, se pelo contrário o modelo corrente não for adequado, $\ell(\hat{\theta})$ deverá ser muito mais pequeno que $\ell(\tilde{\theta})$.

Formalizando, a quantidade

$$D(y, \hat{\theta}) = 2\phi\left(\ell(\tilde{\theta}) - \ell(\hat{\theta})\right) = \phi\sum_{i=1}^{n} D_i, \text{ com } D_i = 2\left\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\right\}$$

é a função de desvio do modelo e a função

$$\frac{D(y, \hat{\theta})}{\phi} = \sum_{i=1}^{n} D_i$$

corresponde ao desvio reduzido.

É função apenas dos dados e decresce à medida que vão sendo adicionadas novas variáveis ao modelo, tomando o valor zero para o modelo saturado. Tem propriedade aditiva para modelos encaixados, pelo que a sua comparação pode ser feita através da diferença entre os seus valores de desvio.
No caso em análise, em que a distribuição da variável resposta é binomial (com n = 1, ie, distribuição de Bernoulli), não se conhece a distribuição do desvio, nem exacta, nem assintótica. Assim a análise do desvio será apenas um guia no estudo da adequabilidade do modelo, através da comparação directa dos valores obtidos para os modelos que se venham a obter e se pretendam comparar, desde que sejam modelos encaixados, uma vez que a diferença do desvio dos dois modelos pode ser interpretada como uma medida da melhoria que os termos adicionais (que se encontram no modelo "maior") trazem em termos de ajustamento à variável resposta.

No caso em que a variável resposta é binária, o desvio não pode ser aproximado por uma distribuição χ^2 (contrariamente ao que sucede com outras distribuições, nomeadamente a binomial em que os dados estão agrupados, ou seja, a variável resposta assume a forma de uma proporção). No entanto, esta aproximação pode geralmente ser usada para a diferença dos desvios de modelos encaixados, sendo por isso muito útil para comparar modelos (e de forma indirecta para testar a inclusão de novos parâmetros a um determinado modelo). A aproximação de um χ^2 é geralmente bastante precisa para a diferença dos desvios, apesar de não o ser para os desvios isoladamente, uma vez que a diferença entre os desvios reduzidos de dois modelos encaixados coincide com a estatística de razão de verosimilhanças, quando a hipótese H_0 diz respeito ao modelo menor e H_1 ao modelo maior.

4.5.3 – Análise dos Resíduos

Os resíduos constituem uma ferramenta de extrema importância para verificar se os pressupostos considerados para a formulação do modelo são ou não correctos. Permitem também identificar observações mal ajustadas, isto é, mal explicadas pelo modelo.

Um resíduo pode ser definido como sendo a expressão da discrepância entre cada valor observado e o respectivo valor ajustado pelo modelo, sendo conveniente usar valores reduzidos e padronizados, isto é, que tenham variância constante unitária.

Existem várias definições de resíduos, sendo os mais frequentemente utilizados os de Pearson, de Anscombe e os resíduos baseados no desvio. Para que se possa concluir que o modelo é adequado e se ajusta bem aos dados, os mesmos devem ser independentes, ter distribuição aproximadamente normal com média 0 e variância constante. Não podem estar relacionados com as variáveis dependentes.
Se alguma destas condições falhar o modelo deve ser revisto, pois isso significa que existe alguma anomalia (a distribuição escolhida não é adequada, a função de ligação pode não ser correcta, podem faltar variáveis independentes, etc…).

A análise dos resíduos é geralmente feita de forma informal com base em representações gráficas. Estas representações gráficas variam consoante o tipo de anomalia cuja presença se pretende detectar (qq-plots, gráficos contra a variável independente e contra as variáveis dependentes, etc..) que permitem aferir se se verificam as condições enumeradas.
No entanto, no caso dos dados binários, os resíduos tomam valores pouco distintos tornando a sua representação gráfica pouco informativa, sendo por isso por vezes usadas algumas transformações dos dados (nomeadamente, dos valores preditos pelo modelo).
A maioria dos estudos sobre esta matéria sugere que no caso binário é mais fiável usar as medidas de diagnóstico anteriormente descritas, nomeadamente o desvio, e não basear a análise da adequabilidade do modelo nos resíduos, que fornecem pouca informação relevante. Mesmo na presença de um modelo correcto, os resíduos poderão estar correlacionados e não terem distribuição normal, motivo pelo qual não se recomenda a sua análise como base de diagnóstico da adequabilidade do modelo.

4.6 – Problemas numéricos

Foram referidos anteriormente alguns problemas numéricos que podem ocorrer ao ajustar um modelo de regressão logística a dados binários. Estes problemas são causados por estruturas específicas de dados que levam a problemas computacionais no cálculo das estimativas dos parâmetros do modelo.
A situação mais simples e óbvia sucede quando os dados podem ser estruturados em tabelas de contingência (ou seja, não são puramente contínuos), nas quais se verificam frequências nulas. Esta situação leva a probabilidades estimadas exactamente iguais a 0 ou 1 e origina problemas de convergência no algoritmo.

Também ocorrem problemas de convergência quando os dados são esparsos ou quando a amostra em estudo tem volume muito reduzido (o que não é aqui o caso). Assim, na prática estas dificuldades podem ocorrer devido ao tamanho da amostra, devido ao número de sucessos / insucessos, devido ao número e ao tipo de variáveis incluídas no modelo, à forma como se relacionam (este tipo de modelo é particularmente sensível à existência de correlação entre as variáveis dependentes), etc…

Estas situações traduzem-se normalmente por valores muito elevados para as estimativas ou para o desvio padrão dos parâmetros do modelo. Pode acontecer obterem-se estimativas para os parâmetros com valores normais, mas a obtenção de valores muito elevados para o desvio padrão deve levar à suspeita de problemas numéricos e deve por isso levar a uma análise cuidada das variáveis incluídas no modelo, da forma como se relacionam entre si, etc…

Estes problemas também podem estar relacionados com a configuração dos dados, que, conforme demonstrado por Albert and Anderson (1984) e Santner and Duffy (1986), pertence sempre a uma de três categorias: separação completa, separação quase completa e sobreposição.
Existe separação completa se houver um vector b que aloca correctamente cada observação ao seu grupo resposta, ie, tal que

$$\begin{cases} bx_i > 0 \text{ para } Y_i = 0 \\ bx_i < 0 \text{ para } Y_i = 1 \end{cases}$$

Se existir alguma observação situada exactamente sobre o vector, os dados dizem-se quase separados. Neste caso tem-se

$$\begin{cases} bx_i \geq 0 \text{ para } Y_i = 0 \\ bx_i \leq 0 \text{ para } Y_i = 1 \end{cases}$$

Destas estruturas de dados (que não ocorrem geralmente com dados contínuos) resultam probabilidades estimadas muito próximas de 0 ou de 1, levando a valores infinitos para as estimativas dos parâmetros e/ou para os seus desvios padrão (a log-verosimilhança tende para 0, levando as estimativas dos parâmetros a crescer para valores infinitos).

Em ambas as situações prova-se que não existem estimadores de máxima verosimilhança, comprometendo a qualidade do modelo resultante do processo iterativo.

No terceiro tipo de estrutura de dados possível, a sobreposição, ao autores dos estudos já mencionados demonstram que os estimadores de máxima verosimilhança existem e são únicos.

Vários autores apresentaram processos bastante complexos para determinar se existe separação completa ou quase completa nos dados em análise. Todos recomendam que, caso não sejam utilizados estes processos, seja sempre avaliada a evolução das estimativas dos parâmetros e respectivos desvios padrão. Se estes tenderem a crescer substancialmente em cada iteração atingindo valores anormalmente elevados, deve concluir-se de que há indícios de separação completa ou quase completa, devendo em seguida tentar-se perceber o motivo para esta estrutura de dados e não se devendo considerar o modelo obtido como válido (são normalmente apresentadas as últimas estimativas obtidas, mas não havendo convergência as mesmas podem ser totalmente incorrectas).

Não existe contudo um processo claro de identificação destas estruturas, nem de como saber qual a variável (ou conjunto de variáveis) que as originam. Geralmente surgem para conjuntos de dados muito reduzidos, para dados considerados esparsos (não sendo apresentada uma definição clara do que são dados esparsos) ou para dados sem variáveis independentes contínuas. Alguns autores referem apenas que a ocorrência deste tipo de dados indica estar-se perante um problema muito complexo, sem apresentar qualquer solução!
Alguns softwares emitem alertas quando a log-verosimilhança se aproxima de 0 ou, quando após algumas iterações, as estimativas ou os seus desvios padrão crescem de forma anormal (o que resulta dos valores da log-verosimilhança). Neste caso só uma análise muito detalhada e cuidada das variáveis incluídas no modelo, da forma como se relacionam entre si e da forma como se relacionam com a variável resposta poderá dar algum indício da origem do problema, e da forma de o resolver.

5 – Formulação do modelo

Foram descritos no capítulo anterior vários métodos que se podem utilizar para escolher as variáveis a incluir no modelo: forward analisys, backwards analisys e e stepwise analisys. Estas três metodologias são bastante distintas, pelo que a comparação dos resultados obtidos com cada uma poderá ter algum interesse.
Como primeira abordagem aplicaram-se os vários métodos descritos à totalidade dos dados disponíveis, obtendo-se sempre os mesmos resultados. O modelo resultante destes processos é o que se apresenta no quadro seguinte:

		Backwards Analisys				
		Tests of Model Effects		Parameter Estimates		
Source	Parameter	Type III		B	Std. Error	Hypothesis Test
		Likelihood Ratio Chi-Square	Sig.			Sig.
(Intercept)	(Intercept)	912.987	0.000	-1.334	0.051	0.000
MetodoPagamento	[MetodoPagamento=Direct Debit]	149,520a	0.000	-0.620	0.053	0.000
	[MetodoPagamento=Others]			0a	.	.
Classificacao	[Classificacao=DadosFixos]	304,941a	0.000	-1.321	0.106	0.000
	[Classificacao=DadosMoveis]			-0.533	0.039	0.000
	[Classificacao=Mista]			-0.895	0.166	0.000
	[Classificacao=Voz]			0a	.	.
Servicos	Servicos	11.508	0.001	-0.102	0.032	0.001
FactUltimoMes	FactUltimoMes	86.434	0.000	-0.004	0.000	0.000
MediaFact6M	MediaFact6M	540.638	0.000	0.028	0.001	0.000
FactMedia	FactMedia	492.773	0.000	-0.028	0.001	0.000
MediaHL	MediaHL	194.766	0.000	3.552	0.257	0.000
HL6M	HL6M	172.320	0.000	-0.420	0.033	0.000
MediaDE	MediaDE	12.950	0.000	2.903	0.808	0.000
DE3M	DE3M	860.363	0.000	3.407	0.130	0.000
Reclamacoes	Reclamacoes	50.896	0.000	0.454	0.061	0.000
Antiguidade	Antiguidade	33.348	0.000	0.000	0.000	0.000
Score	Score	8.044	0.005	0.051	0.018	0.004
	(Scale)			,745b		

a. Set to zero because this parameter is redundant.

b. Computed based on the deviance.

Modelo 1 – Backwards Analysis, conjunto total de contas

Este é um modelo bastante complexo. O elevado número de parâmetros torna a sua interpretação bastante difícil, não permitindo identificar o efeito que cada variável tem sobre o estado da conta.

As variáveis Classificacao e MetodoPagamento definem de certa forma o tipo de cliente, sendo de esperar comportamentos totalmente distintos entre os diferentes tipos de conta.

Sendo estas variáveis categóricas (a variável Classificacao toma quatro valores distintos, a variável MetodoPagamento tem dois valores possíveis), têm de ser transformadas em variáveis binárias para poderem ser introduzidas no modelo. Esta transformação é feita internamente pela aplicação da seguinte forma: uma variável com k valores distintos é transformada em k-1 variáveis, podendo cada uma delas tomar os valores 0 ou 1, consoante a conta pertence à categoria respectiva ou não. São por isso suficientes k-1 variáveis (sabendo se a conta pertence a alguma das k-1 primeiras categorias, sabe-se se pertence à última ou não, tornando por isso a k-ésima variável redundante, conforme indicam os resultados das rotinas SPSS utilizadas).

Uma vez que estas variáveis definem o tipo de conta e que se sabe que existem comportamentos totalmente distintos para contas de tipos diferentes, optou-se pela formulação de modelos para cada subconjunto de dados, segmentados por tipo de conta (ou seja, pela variável Classificacao), seguida de segementação por MetodoPagamento, caso os resultados indiquem que esta variável é relevante, ou seja, caso a formulação de modelos distintos para cada valor desta variável faça também sentido.

Os resultados da análise dos dados segmentados são apresentados em seguida.

5.1 – Segmentação dos dados

O quadro seguinte apresenta as contagens de contas pelas três categorias pelas quais se pretendem segmentar: Estado (variável em estudo), Classificacao e MetodoPagamento:

MetodoPagamento * Estado Conta * Classificacao Crosstabulation

Count				Estado Conta		
				DE - Deactivated	AC - Active	Total
Classificacao	DadosFixos	MetodoPagamento	Direct Debit	5	75	80
			Others	96	1453	1549
		Total		101	1528	1629
	DadosMoveis	MetodoPagamento	Direct Debit	70	958	1028
			Others	2468	9173	11641
		Total		2538	10131	12669
	Mista	MetodoPagamento	Direct Debit	6	125	131
			Others	38	279	317
		Total		44	404	448
	Voz	MetodoPagamento	Direct Debit	355	3566	3921
			Others	2132	7130	9262
		Total		2487	10696	13183

Tal como se tinha verificado na análise inicial, o número de contas difere muito de umas categorias para outras. Sendo também de esperar que tenham características completamente diferentes, faz sentido proceder ao seu estudo de forma separada para averiguar se de facto assim é.

5.2 – Contas de Voz

Prosseguindo com a análise baseada na separação das contas por tipo, consideraram-se apenas as contas de Voz (isto é, que têm apenas serviços de voz, num total de 13.183) como base para a formulação do modelo, usando as técnicas anteriormente descritas (backwards, forward e stepwise analisys). O resultado obtido foi o mesmo para as três metodologias e é apresentado no quadro seguinte:

Source	Parameter	Tests of Model Effects		Parameter Estimates		
		Type III		B	Std. Error	Hypothesis Test
		Likelihood Ratio Chi-Square	Sig.			Sig.
(Intercept)	(Intercept)	1723,996	,000	-1,964	,0520	,000
MetodoPagamento	[MetodoPagamento=Direct Debit]	72,654	,000	-,517	,0621	,000
	[MetodoPagamento=Others]			0a	.	.
FactUltimoMes	FactUltimoMes	93,166	,000	-,005	,0005	,000
MediaFact3M	MediaFact3M	20,257	,000	,005	,0011	,000
MediaFact6M	MediaFact6M	132,900	,000	,021	,0019	,000
FactMedia	FactMedia	317,373	,000	-,027	,0017	,000
MediaHL	MediaHL	237,342	,000	5,091	,3485	,000
HL6M	HL6M	189,459	,000	-,592	,0473	,000
DE3M	DE3M	805,124	,000	3,553	,1565	,000
Antiguidade	Antiguidade	97,774	,000	,000	1,6506E-5	,000
Score	Score	296,659	,000	,506	,0294	,000
	(Scale)			,757b		

a. Set to zero because this parameter is redundant.

b. Computed based on the deviance.

Modelo 2 – Todos os métodos, Contas de Voz

Este modelo, embora mais simples que o obtido para todas as contas, continua a ser um modelo algo complexo devido ao número de variáveis incluídas. Permite verificar que para as contas de voz os factores mais relevantes estão maioritariamente relacionados com facturação e pagamentos. Estão incluídas todas as variáveis de facturação (médias e últimos valores), bem como variáveis relacionadas com atrasos nos pagamentos (MediaHL, HL6M e DE3M). Além das questões directamente relacionadas com facturação e pagamentos, são também relevantes as variáveis Antiguidade e Score.
Uma vez que a variável MetodoPagamento é relevante, e considerando o facto de esta variável só poder tomar dois valores distintos, dividindo as contas em dois grupos que se esperam ter características distintas, procedeu-se à modelação separada das contas que têm método de pagamento por débito directo e das restantes.

Para as contas sem débito directo (MetodoPagamento = Others, n = 9262), obtiveram-se resultados distintos consoante a estratégia de modelação escolhida.

O método backwards analisys conduziu à exclusão das seguintes variáveis:

Backwards Analisys	
Iteration 1	Exclusão da variável DE6M
Iteration 2	Exclusão da variável HL3M
Iteration 3	Exclusão da variável MediaDE
Iteration 4	Exclusão da variável Servicos
Iteration 5	Exclusão da variável Reclamacoes
Iteration 6	Exclusão da variável DistritoAgrupIlhas

O modelo resultante, cujo detalhe é apresentado no quadro seguinte, tem as mesmas variáveis que o modelo obtido para todas as contas de voz, o que não surpreende uma vez que as contas sem débito directo têm um peso significativo neste universo (correspondem a 70% do total).

		Iteration 7				
		Tests of Model Effects		Parameter Estimates		
Source	Parameter	Type III		B	Std. Error	Hypothesis Test
		Likelihood Ratio Chi-Square	Sig.			Sig.
(Intercept)	(Intercept)	1379,934	,000	-1,947	,0588	,000
FactUltimoMes	FactUltimoMes	53,732	,000	-,004	,0008	,000
MediaFact3M	MediaFact3M	17,142	,000	,005	,0017	,001
MediaFact6M	MediaFact6M	78,679	,000	,019	,0023	,000
FactMedia	FactMedia	209,795	,000	-,025	,0020	,000
MediaHL	MediaHL	219,960	,000	5,284	,3780	,000
HL6M	HL6M	178,371	,000	-,619	,0513	,000
DE3M	DE3M	704,695	,000	3,561	,1680	,000
Antiguidade	Antiguidade	87,873	,000	,000	2,0329E-5	,000
Score	Score	217,524	,000	,502	,0343	,000
	(Scale)			,835		

Modelo 3.1 – Backwards Analysis, Contas de Voz sem Débito Directo

Pelo método de stepwise / forward analisys, obtém-se um modelo ligeiramente diferente:

Forward Analisys	
Iteration 1	Inclusão da variável DE3M
Iteration 2	Inclusão da variável Score
Iteration 3	Inclusão da variável Antiguidade
Iteration 4	Inclusão da variável MediaDE
Iteration 5	Inclusão da variável MediaHL
Iteration 6	Inclusão da variável HL6M, Exclusão da variável MediaDE
Iteration 7	Inclusão da variável Servicos

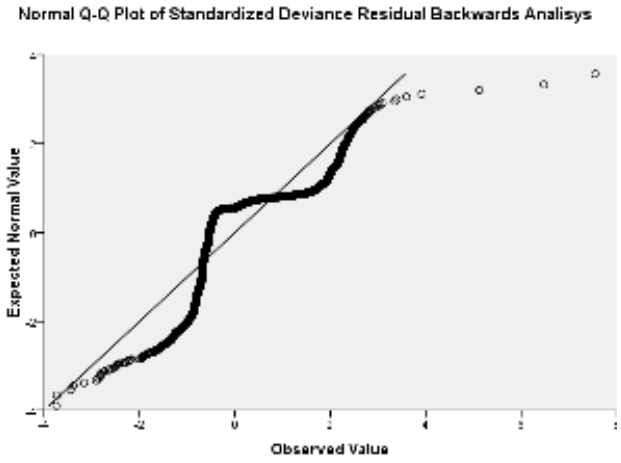
Teste à inclusão de novas variáveis	
DistritoAgrupIlhas	0.06
FactUltimoMes	0.185
MediaFact3M	0.128
MediaFact6M	0.214
FactMedia	0.085
HL3M	0.335
MediaDE	0.13
DE6M	0.717
Reclamacoes	0.095

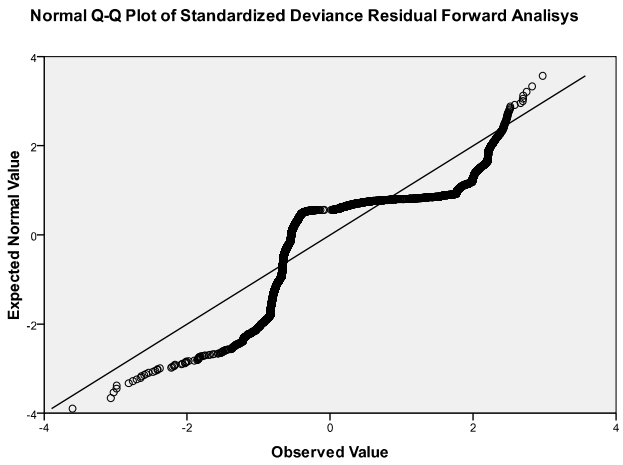
		Iteration 8				
		Tests of Model Effects		Parameter Estimates		
Source	Parameter	Type III		B	Std. Error	Hypothesis Test
		Likelihood Ratio Chi-Square	Sig.			Sig.
(Intercept)	(Intercept)	662,340	,000	-1,871	,0699	,000
DE3M	DE3M	769,547	,000	3,540	,1602	,000
Score	Score	217,643	,000	,468	,0321	,000
Antiguidade	Antiguidade	107,924	,000	,000	2,0696E-5	,000
MediaHL	MediaHL	241,279	,000	5,325	,3694	,000
HL6M	HL6M	201,775	,000	-,636	,0504	,000
Servicos	Servicos	7,174	,007	-,103	,0400	,010
	(Scale)			,855		

Modelo 3.2 – Forward / Stepwise Analysis, Contas de Voz sem Débito Directo

Este modelo é mais simples (por ter um menor número de variáveis). São consideradas relevantes as mesmas variáveis, excepto todas as relacionadas com facturação (4 variáveis). Em contrapartida este modelo inclui o número de serviços associados à conta (variável excluída do primeiro modelo).

A análise dos resíduos resultantes de ambos os modelos é inconclusiva. Tal como se referiu anteriormente, no caso em que a variável resposta é do tipo binário, os resíduos fornecem pouca informação relevante, pelo que são pouco úteis para a validação do modelo.
Apesar de, tal como era expectável, os resíduos não se ajustarem bem à distribuição normal, o desvio é aparentemente inferior no segundo modelo. Além disso, sendo este mais simples (do ponto de vista de número de variáveis envolvidas), considera-se este como mais adequado aos dados em análise.





Fazendo a mesma análise para as contas com método de pagamento por débito directo (n = 3921), obtém-se o mesmo modelo, independentemente da estratégia de selecção de variáveis utilizada:

		Tests of Model Effects		Parameter Estimates		
Source	**Parameter**	Type III		B	Std. Error	Hypothesis Test
		Likelihood Ratio Chi-Square	Sig.			Sig.
(Intercept)	(Intercept)	519,031	,000	-2,667	,1340	,000
Antiguidade	Antiguidade	6,494	,011	-7,588E-5	2,9534E-5	,010
DE3M	DE3M	15,178	,000	2,978	,9234	,001
FactMedia	FactMedia	132,337	,000	-,037	,0037	,000
FactUltimoMes	FactUltimoMes	34,102	,000	-,005	,0012	,000
HL6M	HL6M	10,630	,001	,550	,1557	,000
MediaFact6M	MediaFact6M	111,060	,000	,033	,0033	,000
Score	Score	77,450	,000	,561	,0615	,000
	(Scale)			,569		

Modelo 4.1 – Todos os métodos, Contas de Voz com Débito Directo

Durante o processo de selecção do modelo, quer partindo do mais simples quer partindo do mais complexo, surgiram problemas de convergência com a variável Distrito. Em todos os modelos que incluam esta variável, é dado o seguinte alerta, indicativo de problemas de convergência:

Warnings

A quasi-complete separation may exist in the data. The maximum likelihood estimates do not exist.

The GENLIN procedure continues despite the above warning(s). Subsequent results shown are based on the last iteration. Validity of the model fit is uncertain.

Os resultados detalhados apresentam indícios deste problema: para a variável dicotómica associada à categoria Portalegre, o desvio padrão tem valores anormalmente elevados:

Parameter Estimates

Parameter	B	Std. Error	95% Profile Likelihood Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-2,289	,3157	-2,945	-1,700	52,583	1	,000
Antiguidade	-6,931E-5	2,9724E-5	,000	-1,070E-5	5,437	1	,020
DE3M	2,978	,9330	1,348	5,164	10,185	1	,001
FactMedia	-,037	,0036	-,045	-,030	103,606	1	,000
FactUltimoMes	-,005	,0012	-,008	-,003	17,268	1	,000
HL6M	,566	,1566	,248	,871	13,087	1	,000
MediaFact6M	,033	,0033	,027	,040	100,175	1	,000
Score	,574	,0621	,452	,696	85,441	1	,000
[DistritoAgrupIlhas=Aveiro]	-,159	,3388	-,803	,534	,220	1	,639
[DistritoAgrupIlhas=Beja]	-,029	,5608	-1,212	1,024	,003	1	,958
[DistritoAgrupIlhas=Braga]	-,173	,3444	-,830	,529	,253	1	,615
[DistritoAgrupIlhas=Braganca]	-,413	,5380	-1,554	,594	,590	1	,442
[DistritoAgrupIlhas=CasteloBranco]	-1,554	,7840	-3,389	-,199	3,930	1	,047
[DistritoAgrupIlhas=Coimbra]	-,087	,3404	-,734	,609	,065	1	,799
[DistritoAgrupIlhas=Evora]	-,098	,4976	-1,126	,852	,039	1	,844
[DistritoAgrupIlhas=Faro]	-,286	,3529	-,962	,430	,655	1	,418
[DistritoAgrupIlhas=Guarda]	-,170	,5467	-1,328	,854	,097	1	,756
[DistritoAgrupIlhas=Ilhas]	-,185	,3996	-,971	,608	,213	1	,644
[DistritoAgrupIlhas=Leiria]	-,240	,3435	-,895	,460	,488	1	,485
[DistritoAgrupIlhas=Lisboa]	-,501	,2998	-1,058	,126	2,799	1	,094
[DistritoAgrupIlhas=Portalegre]	-20,464	12431,0817	.a	.a	,000	1	,999
[DistritoAgrupIlhas=Porto]	-,525	,3123	-1,109	,123	2,824	1	,093
[DistritoAgrupIlhas=Santarem]	-,660	,3590	-1,352	,066	3,378	1	,066
[DistritoAgrupIlhas=Setubal]	-,524	,3302	-1,149	,154	2,516	1	,113
[DistritoAgrupIlhas=VianaDoCastelo]	-1,227	,6234	-2,614	-,100	3,877	1	,049
[DistritoAgrupIlhas=VilaReal]	-,630	,5557	-1,807	,408	1,285	1	,257
	0b	.	.	.	.	.	.
(Scale)	,567c						

Dependent Variable: Estado Conta

Model: (Intercept), Antiguidade, DE3M, FactMedia, FactUltimoMes, HL6M, MediaFact6M, Score, DistritoAgrupIlhas

a. Unable to compute because some convergence criteria were not satisfied.

b. Set to zero because this parameter is redundant.

c. Computed based on the deviance.

Modelo 4.2 –Contas de Voz com Débito Directo, modelo incluindo a variável Distrito

Uma análise detalhada aos dados classificados por distrito permite concluir que o problema está relacionado com o facto de não haver contas deste tipo (apenas com serviços de Voz e com método de pagamento por débito directo) desactivas em Portalegre:

DistritoAgrupIlhas * Estado Conta Crosstabulation

Count		Estado Conta		Total
		DE - Deactivated	AC - Active	
DistritoAgrupIlhas	Aveiro	22	180	202
	Beja	3	19	22
	Braga	20	153	173
	Braganca	4	37	41
	CasteloBranco	2	31	33
	Coimbra	22	164	186
	Evora	4	32	36
	Faro	17	154	171
	Guarda	3	27	30
	Ilhas	9	73	82
	Leiria	20	176	196
	Lisboa	122	1349	1471
	Portalegre	0	22	22
	Porto	54	541	595
	Santarem	14	180	194
	Setubal	26	296	322
	VianaDoCastelo	2	42	44
	VilaReal	3	36	39
	Viseu	8	54	62
Total		355	3566	3921

O facto de não haver contas desactivas numa das categorias é a causa dos problemas de convergência. Poderia daqui tirar-se a conclusão de que todas as contas deste distrito se mantêm activas (ou seja, que a probabilidade de desactivar para as contas pertencentes a este distrito é nula), mas esta conclusão não está obviamente correcta.

Pode-se considerar que o facto de termos muito mais contas activas que desactivas e frequências muito baixas para determinados distritos (e métodos de pagamento) conduziram a frequências nulas, o que não pode levar por si só a nenhuma inferência sobre a probabilidade de uma conta se manter activa ou não. Esta situação está normalmente associada a uma coincidência e não descreve geralmente um fenómeno ou uma conclusão importante acerca dos dados em análise.

Não considerando as contas de Portalegre, os modelos obtidos são exactamente os mesmos, pelo que se conclui que a variável Distrito não influencia de forma significativa o estado da conta.

5.3 – Contas de Dados Móveis

Tal como se fez para as contas de voz, a análise às contas de dados móveis (n = 12.669) consistiu numa primeira fase na formulação de um modelo com todas as variáveis disponíveis, aplicando as várias metodologias para seleccionar as que são significativas. Também aqui se chegou sempre ao mesmo modelo, com os seguintes parâmetros:

Source	Parameter	Tests of Model Effects		Parameter Estimates		
		Type III		B	Std. Error	Hypothesis Test
		Likelihood Ratio Chi-Square	Sig.			Sig.
(Intercept)	(Intercept)	1111,105	,000	-1,526	,0509	,000
MetodoPagamento	[MetodoPagamento=Direct Debit]	90,562	,000	-1,012	,1203	,000
	[MetodoPagamento=Others]			0a	.	.
FactUltimoMes	FactUltimoMes	11,437	,001	-,002	,0006	,002
MediaFact3M	MediaFact3M	9,127	,003	-,004	,0013	,002
MediaFact6M	MediaFact6M	242,327	,000	,048	,0035	,000
FactMedia	FactMedia	410,426	,000	-,069	,0040	,000
MediaHL	MediaHL	34,940	,000	2,251	,3977	,000
HL6M	HL6M	43,084	,000	-,326	,0527	,000
MediaDE	MediaDE	92,572	,000	12,650	1,3386	,000
DE3M	DE3M	579,027	,000	4,296	,2080	,000
DE6M	DE6M	38,653	,000	-1,236	,2002	,000
Reclamacoes	Reclamacoes	59,919	,000	,807	,1033	,000
Antiguidade	Antiguidade	128,075	,000	,001	,0001	,000
Score	Score	34,298	,000	-,161	,0278	,000
	(Scale)			,728b		

a. Set to zero because this parameter is redundant.

b. Computed based on the deviance.

Modelo 5 – Todos os métodos, Contas de Dados Móveis

À semelhança do que aconteceu com as contas de voz, os dados foram divididos de acordo com o método de pagamento (uma vez que esta variável foi considerada significativa). No caso das contas que não pagam as suas facturas por débito directo (MetodoPagamento = Others, n = 11.641), obtiveram-se os seguintes resultados:

Source	Parameter	Tests of Model Effects		Parameter Estimates		
		Type III		B	Std. Error	Hypothesis Test
		Likelihood Ratio Chi-Square	Sig.			Sig.
(Intercept)	(Intercept)	934,859	,000	-1,467	,0522	,000
Antiguidade	Antiguidade	96,282	,000	,001	,0001	,000
DE3M	DE3M	540,442	,000	4,297	,2146	,000
DE6M	DE6M	38,139	,000	-1,266	,2063	,000
FactMedia	FactMedia	396,460	,000	-,070	,0042	,000
FactUltimoMes	FactUltimoMes	14,095	,000	-,002	,0007	,002
HL6M	HL6M	35,783	,000	-,308	,0549	,000
MediaDE	MediaDE	95,154	,000	13,265	1,3847	,000
MediaFact3M	MediaFact3M	8,055	,005	-,004	,0013	,003
MediaFact6M	MediaFact6M	234,291	,000	,049	,0036	,000
MediaHL	MediaHL	28,431	,000	2,097	,4139	,000
Reclamacoes	Reclamacoes	56,658	,000	,807	,1060	,000
Score	Score	38,191	,000	-,175	,0287	,000
	(Scale)			,754		

Modelo 6 – Todos os métodos, Contas de Dados Móveis sem Débito Directo

Para além das variáveis relacionadas com a facturação e os prazos de pagamento (nº de hotlines e desactivações associadas aos pagamentos fora de prazo), são também relevantes a antiguidade, o

score e as reclamações – variável não significativa no caso da voz, independentemente do método de pagamento.

Para as contas com débito directo (n = 1028), obtiveram-se dois modelos diferentes. Partindo do modelo mais complexo (e eliminando a variável menos significativa em cada iteração), obtiveram-se os resultados seguintes:

		Tests of Model Effects		Parameter Estimates		
Source	Parameter	Type III		B	Std. Error	Hypothesis Test
		Likelihood Ratio Chi-Square	Sig.			Sig.
(Intercept)	(Intercept)	497,969	,000	-4,339	,2501	,000
FactUltimoMes	FactUltimoMes	6,664	,010	,007	,0026	,012
MediaFact3M	MediaFact3M	4,948	,026	-,018	,0082	,029
MediaFact6M	MediaFact6M	7,765	,005	,037	,0142	,009
FactMedia	FactMedia	5,254	,022	-,024	,0117	,042
MediaHL	MediaHL	5,115	,024	2,789	1,2111	,021
HL3M	HL3M	16,177	,000	,903	,2289	,000
HL6M	HL6M	7,324	,007	-,474	,1752	,007
MediaDE	MediaDE	17,535	,000	12,580	3,0486	,000
Reclamacoes	Reclamacoes	6,570	,011	1,414	,4902	,004
Antiguidade	Antiguidade	61,794	,000	,002	,0003	,000
Score	Score	6,181	,013	,316	,1261	,012
	(Scale)			,415		

Modelo 7.1 – Backwards Analisys, Contas de Dados Móveis com Débito Directo

Partindo do modelo mais simples e adicionando em cada iteração a variável mais significativa (controlando em todas as fases do processo a significância das variáveis já incluídas, bem como o valor do desvio associado a cada modelo), obteve-se o seguinte:

		Tests of Model Effects		Parameter Estimates		
Source	Parameter	Type III		B	Std. Error	Hypothesis Test
		Likelihood Ratio Chi-Square	Sig.			Sig.
(Intercept)	(Intercept)	730,120a	,000	-4,361	,2374	,000
DE3M	DE3M	44,986	,000	5,146	,9305	,000
Antiguidade	Antiguidade	58,898	,000	,002	,0003	,000
Reclamacoes	Reclamacoes	10,139	,001	1,631	,4509	,000
MediaHL	MediaHL	25,517	,000	6,797	1,2737	,000
HL6M	HL6M	14,965	,000	-,691	,1744	,000
FactUltimoMes	FactUltimoMes	8,064	,005	,007	,0026	,006
MediaDE	MediaDE	5,913	,015	-12,454	5,3664	,020
Score	Score	3,968	,047	,241	,1203	,045
(Scale)	(Scale)			,405		

Modelo 7.2 – Forward / Stepwise Analisys, Contas de Dados Móveis com Débito Directo

O segundo modelo é bastante mais simples que o primeiro, não considerando as variáveis de facturação (à excepção da facturação do último mês) nem o número de Hotlines dos últimos 3 meses. Considera em contrapartida o número de desactivações dos últimos 3 meses.

Uma vez que a interpretação dos parâmetros é muito semelhante, faz sentido considerar o segundo modelo como mais adequado, visto ser bastante mais parcimonioso.
Não existem diferenças significativas entre as contas com e sem débito directo, o tipo de variáveis consideradas (isto é, o tipo de informação a que dizem respeito) são semelhantes.

Ocorreram problemas de convergência em todos os modelos com a variável Distrito. O mesmo sucedeu quando nos modelos que tinham a variável DE3M em conjunto com a variável Distrito ou a variável DE6M, tendo sido emitido no output o aviso de separação quase completa dos dados.

DistritoAgrupIlhas * DE3M Crosstabulation

Count		DE3M 0	DE3M 1	Total
DistritoAgrupIlhas	Aveiro	74	1	75
	Beja	18	0	18
	Braga	75	0	75
	Braganca	18	0	18
	CasteloBranco	12	0	12
	Coimbra	53	2	55
	Evora	9	2	11
	Faro	57	3	60
	Guarda	10	0	10
	Ilhas	21	1	22
	Leiria	75	0	75
	Lisboa	238	2	240
	Portalegre	8	0	8
	Porto	139	3	142
	Santarem	60	1	61
	Setubal	70	3	73
	VianaDoCastelo	28	0	28
	VilaReal	17	0	17
	Viseu	27	1	28
Total		1009	19	1028

Analisando os dados associados às categorias cruzadas destas variáveis, conclui-se que o problema está novamente relacionado com contagens nulas, que levam a probabilidades estimadas de 0 (ou 1) e consequentemente a parâmetros a tender para valores infinitos. Estas variáveis não puderam por este motivo ser consideradas simultaneamente no modelo (tendo sido a variável Distrito excluída logo de início, uma vez que havia categorias com contagens nulas para o estado desactivo, e que nos modelos formulados considerando esta variável sem estas categorias a mesma nunca foi considerada significativa).

DE6M * DE3M Crosstabulation

Count		DE3M 0	DE3M 1	
DE6M	0	1000	0	1000
	1	9	19	28
Total		1009	19	1028

5.4 – Contas de Dados Fixos

A análise às contas de dados fixos (total de 1649) produziu dois modelos muito semelhantes, havendo diferença apenas na variável referente ao número de desactivações incluída: um dos modelos considerou relevante o número de desactivações dos últimos 3 meses, o outro a 6. Estas variáveis são muito semelhantes pelo que tanto uma como outra pode ser considerada. As medidas de diagnóstico são muito semelhantes, pelo que ambos os modelos são adequados. Apresentam-se os parâmetros de um deles:

		Tests of Model Effects		Parameter Estimates		
Source	Parameter	Type III		B	Std. Error	Hypothesis Test
		Likelihood Ratio Chi-Square	Sig.			Sig.
MetodoPagamento	[MetodoPagamento=Direct Debit]	15,823	,000	,240	,3934	,542
	[MetodoPagamento=Others]			-,588	,1845	,001
Servicos	Servicos	127,657	,000	-1,084	,1042	,000
MediaFact6M	MediaFact6M	136,059	,000	,174	,0200	,000
FactMedia	FactMedia	111,644	,000	-,181	,0221	,000
HL3M	HL3M	56,533	,000	7,555	1,4701	,000
HL6M	HL6M	43,048	,000	-6,197	1,4499	,000
DE6M	DE6M	23,320	,000	3,346	,8645	,000
Score	Score	32,501	,000	-,297	,0573	,000
Reclamacoes	Reclamacoes	3,740	,053	,747	,3451	,030
	(Scale)			,380		

Modelo 8 – Backwards Analisys, Contas de Dados Fixos

No processo de escolha das variáveis ocorreram novamente problemas de convergência com o Distrito e o número de desactivações (dos últimos 3 e dos últimos 6 meses), tendo a sua análise mostrado que a causa é a que já anteriormente se identificou: categorias com contagens nulas para contas com estado desactivo.

Sendo a variável Método de Pagamento significativa, à semelhança do que foi feito nas categorias de contas anteriores, analisaram-se separadamente as contas com e sem débito directo.

Para as contas sem débito directo (n = 1549), obteve-se o modelo seguinte:

		Tests of Model Effects		Parameter Estimates		
Source	Parameter	Type III		B	Std. Error	Hypothesis Test
		Likelihood Ratio Chi-Square	Sig.			Sig.
(Intercept)	(Intercept)	12,230	,000	-,653	,1883	,001
Servicos	Servicos	114,878	,000	-1,049	,1062	,000
MediaFact6M	MediaFact6M	139,036	,000	,185	,0211	,000
FactMedia	FactMedia	117,209	,000	-,195	,0234	,000
HL3M	HL3M	56,009a	,000	7,587	1,4832	,000
HL6M	HL6M	42,876	,000	-6,255	1,4658	,000
DE6M	DE6M	23,984	,000	3,430	,8821	,000
Reclamacoes	Reclamacoes	4,040	,045	,790	,3507	,024
Score	Score	25,862	,000	-,274	,0591	,000
	(Scale)			,381		

Modelo 9 – Backwards Analisys, Contas de Dados Fixos sem Débito Directo

Este modelo é muito semelhante ao que se obteve para todas contas de dados fixos, o que seria de esperar uma vez que a categoria em estudo (MetodoPagamento = Others) constitui a grande maioria dos dados.

Já no caso das contas com débito directo (presentes em número muito reduzido, n = 80), o modelo obtido é bastante mais simples:

Source	Parameter	Tests of Model Effects		Parameter Estimates		
		Type III		B	Std. Error	Hypothesis Test
		Likelihood Ratio Chi-Square	Sig.			Sig.
Servicos	Servicos	44,586	,000	-1,315	,2580	,000
FactUltimoMes	FactUltimoMes	12,120	,001	-,045	,0197	,022
FactMedia	FactMedia	19,139	,000	,101	,0231	,000
Score	Score	5,005	,028	-,411	,1899	,030
	(Scale)			,270		

Modelo 10 – Backwards Analisys, Contas de Dados Fixos com Débito Directo

Ocorreram problemas numéricos com um elevado número de variáveis, todos relacionados com o número reduzido de contas em análise: nuns casos por se verificarem contagens nulas num dos estados, noutro porque a variável toma apenas dois valores possíveis (por exemplo, só há contas com 0 ou 1 desactivação), ocorrendo por isso o seguinte erro: "*Unable to compute because the estimable function has zero degrees of freedom."*
As variáveis para as quais se verificaram estes problemas foram excluídas do modelo.

5.5 – Contas Mistas

A análise da última categoria, com 448 contas, conduziu ao modelo seguinte:

Source	Parameter	Tests of Model Effects		Parameter Estimates		
		Type III		B	Std. Error	Hypothesis Test
		Likelihood Ratio Chi-Square	Sig.			Sig.
(Intercept)	(Intercept)	9,619	,002	-1,321	,4964	,008
MetodoPagamento	[MetodoPagamento=Direct Debit]	5,214	0.022	-,774	,3592	,031
	[MetodoPagamento=Others]			0a	.	.
Servicos	Servicos	10,368	,001	-,562	,2033	,006
FactUltimoMes	FactUltimoMes	14,619	,000	-,011	,0033	,001
MediaFact6M	MediaFact6M	18,608	,000	,013	,0029	,000
HL6M	HL6M	3,619	,057	,210	,1018	,040
DE3M	DE3M	27,386	,000	2,476	,4982	,000
Reclamacoes	Reclamacoes	5,984	,014	,646	,2425	,008
	(Scale)			,512b		

a. Set to zero because this parameter is redundant.

b. Computed based on the deviance.

Modelo 11 – Backwards Analisys, Contas Mistas

Não se verificaram diferenças significativas entre os modelos obtidos utilizando cada um dos métodos – diferiram apenas numa variável (um dos modelos considerou a variável HL6M, o outro MediaHL. Sendo estas variáveis correlacionadas e baseadas no mesmo tipo de eventos, não se considera esta diferença relevante).

Destas contas, 317 não têm débito directo. Da sua análise resultou um modelo com os parâmetros seguintes (iguais para as duas estratégias de modelação seguidas):

		Tests of Model Effects		Parameter Estimates		
Source	Parameter	Type III		B	Std. Error	Hypothesis Test
		Likelihood Ratio Chi-Square	Sig.			Sig.
Servicos	Servicos	152,228	,000	-1,004	,1038	,000
FactUltimoMes	FactUltimoMes	8,802	,003	-,009	,0034	,008
DE3M	DE3M	62,930	,000	3,668	,5133	,000
Reclamacoes	Reclamacoes	4,993	,025	,640	,2656	,016
Antiguidade	Antiguidade	6,868	,009	,000	,0001	,005
	(Scale)			,582		

Modelo 12 –Todos os Métodos, Contas Mistas sem Débito Directo

Finalmente, as restantes 131 contas com pagamento por débito directo levaram à formulação do modelo seguinte:

		Tests of Model Effects		Parameter Estimates		
Source	Parameter	Type III		B	Std. Error	Hypothesis Test
		Likelihood Ratio Chi-Square	Sig.			Sig.
(Intercept)	(Intercept)	9,104	,003	-2,377	,6943	,001
MediaFact6M	MediaFact6M	30,417	,000	,023	,0051	,000
Servicos	Servicos	7,804	,005	-,891	,3562	,012
FactUltimoMes	FactUltimoMes	3,738	,053	-,009	,0048	,074
	(Scale)			,302		

Modelo 13 –Forward Analisys, Contas Mistas com Débito Directo

Também neste caso não se verificaram diferenças significativas entre os modelos obtidos usando Backwards e Forward / Stepwise Analisys.
Ocorreram mais uma vez problemas numéricos com a inclusão de parte das variáveis, todos relacionados com a estrutura de dados de cada uma delas (em todos os casos foram identificadas contagens nulas nas respectivas tabelas de contingência com dados cruzados por estado da conta), tendo por esse motivo sido excluídas.

6- Conclusões

A análise efectuada aos dados disponíveis permite concluir que existem diferenças no que diz respeito às variáveis que se podem considerar relevantes em termos do estado da conta, consoante a categoria a que esta pertence. O quadro seguinte resume as variáveis consideradas significativas em cada caso (agrupadas pelos tipos de dados que representam):

		Metodo Pagamento		
		Todos	Others	DirectDebit
Tipo de Conta	DadosFixos	Metodo de Pagamento		
		Servicos	Servicos	Servicos
		Facturação	Facturação	Facturação
		Score	Score	Score
		Hotlines	Hotlines	
		Desactivações	Desactivações	
		Reclamacoes	Reclamacoes	
	DadosMoveis	Metodo de Pagamento		
		Facturação	Facturação	Facturação
		Hotlines	Hotlines	Hotlines
		Desactivações	Desactivações	Desactivações
		Reclamacoes	Reclamacoes	Reclamacoes
		Antiguidade	Antiguidade	Antiguidade
		Score	Score	Score
	Mista	Metodo de Pagamento		
		Servicos	Servicos	Servicos
		Facturação	Facturação	Facturação
		Desactivações	Desactivações	
		Reclamacoes	Reclamacoes	
		Hotlines	Antiguidade	
	Voz	Metodo de Pagamento		
		Facturação		Facturação
		Hotlines	Hotlines	Hotlines
		Desactivações	Desactivações	Desactivações
		Antiguidade	Antiguidade	Antiguidade
		Score	Score	Score
			Servicos	

Começando por identificar as principais conclusões que se podem retirar da comparação dos quatro tipos de contas, podem-se salientar as seguintes:

- As variáveis relativas ao método de pagamento, à facturação, ao número de hotlines e ao número de desactivações são significativas, independentemente do tipo de conta considerado.
- O número de reclamações é relevante em todas as categorias excepto na de voz
- A antiguidade e o score são factores relevantes quer no caso das contas de voz, quer no caso das contas de dados móveis
- O número de serviços associados à conta tem importância para as contas de dados fixos e mistas.

No que diz respeito à comparação entre contas com e sem débito directo, também se podem fazer observações interessantes:

- No caso dos dados móveis não há diferenças consoante o método de pagamento, os modelos obtidos têm variáveis semelhantes
- Para as contas de voz, a facturação só é relevante para as contas com débito directo. O modelo para as restantes contas não considera esta variável, mas inclui em contrapartida o nº de serviços.
- Para as contas mistas e de dados fixos obtiveram-se modelos bastante mais simples no caso para método de pagamento por débito directo, mas tal pode dever-se ao volume reduzido de dados.

Todos os resultados foram obtidos através da formulação de modelos com variável resposta do tipo binário (0 ou 1). A utilização destes dados levantou alguns problemas: muitos resultados (nomeadamente os relacionados com as qualidades assintóticas do desvio) só são válidos para dados agrupados, isto é, dados em que a variável resposta corresponde a uma proporção (ou a uma probabilidade de sucesso) – resposta do tipo binomial. A utilização deste tipo de resposta tem também algumas vantagens na interpretação dos resíduos obtidos.
O Teorema do Limite Central mostra que a distribuição binomial pode ser aproximada por uma distribuição normal, o que leva alguns autores a aconselhar o uso de modelos com este tipo de resposta – apesar de, conforme se demonstrou ao longo deste trabalho, os modelos lineares generalizados serem válidos para respostas do tipo de binário.
Uma vez que nem todas as variáveis independentes são do tipo categórico, o agrupamento de dados – com o objectivo de se trabalhar com uma variável de resposta binomial, ou seja com uma probabilidade de sucesso - implicaria criar categorias / intervalos para todas as variáveis contínuas (com variáveis respostas do tipo contínuo não há observações com valores exactamente iguais para cada uma das variáveis, ou seja, não é possível agrupar directamente as observações).
A criação de categorias para as variáveis contínuas introduziria necessariamente algum erro – a discretização das variáveis implica perda de precisão. Além disso a escolha dos intervalos teria de ser feita por forma a obter-se um conjunto de observações comuns com algum peso (por forma a que as probabilidades de sucesso ou insucesso respectivas tivessem algum significado), o que poderia obrigar à criação de intervalos com alguma amplitude e mais uma vez perda de precisão.
Por este motivo optou-se pela utilização dos dados disponíveis sem transformações, ou seja com respostas binárias, sem agrupamento de observações.
Esta decisão teve algumas implicações e esteve possivelmente na base de algumas dificuldades, nomeadamente os problemas de convergência, pelo que seria interessante formular os vários modelos com base em dados agrupados e comparar os resultados. Também seria uma oportunidade de avaliar se os problemas numéricos surgidos em vários modelos se manteriam ou não (alguns autores apontam o facto de os dados estarem não agrupados como uma possível causa para a não convergência dos algoritmos).

Também é interessante notar que as diferenças entre as várias estratégias de modelação (backwards, forward e stepwise analisys) foram pouco relevantes, estas conduziram quase sempre ao mesmo resultado ou, no caso de resultados distintos, a modelos muito próximos, com variáveis do mesmo tipo. Parece assim não ser muito importante seguir por um tipo de algoritmo ou por outro com este tipo de dados, também aqui teria interesse averiguar se o mesmo acontece com os dados agrupados (isto é com resposta do tipo binomial).

Bibliografia

Albert, A. , and Anderson, J. A. (1984), *On the existence of maximum likelihood estimates in logistic regression models*, Biometrika, 71 , 1-10

Amaral Turkman, M. A. e Silva, G. L. (2000), *Modelos Lineares Generalizados – da teoria à prática*, VIII congresso anual da Sociedade Portuguesa de Estatística

Collett, David (2002), *Modelling Binary Data, Second Edition*, Chapman and Hall, London

Dobson, A.J. (1991), *An introduction to Generalized Linear Models*, Chapman and Hall, London

Hosmer D. and Lemeshow S. (2000*), Applied Logistic Regression, Second Edition*, Wiley Series in Probability and Statistics

McCullagh P. and Nelder J.A. (1989), *Generalized Linear Models, Second Edition*, Chapman and Hall, London

Santner, T. J., and Duffy, D. E. (1986). *A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models*. Biometrika 73, 755-758

Silvapulle, M. J., (1981). *On the existence of maximum likelihood estimators for the binomial response models*. Journal of the Royal Statistical Society, Series B 43, 310-313

Silvapulle, M. J., and Burridge, J. (1986). *Existence of maximum likelihood estimates in regression models for grouped and ungrouped data*. Journal of the Royal Statistical Society, Series B 48, 100-106

So, Ying, (1993) *A Tutorial on Logistic Regression*, SAS Institute Inc., Cary, NC

Ventura L. (2004-2005*)*, *Modelli Lineari Generalizzati, Modelli Statistici II*, Dipartimento di Scienze Statistiche, Universit`a degli Studi di Padova

Webb, Wilson and Williams (2002), *An Examination of Statistical Software Packages for Binary Logistic Models on Quasi-complete: Applications to Alcohol Abuse Data*, Joint Statistical Meetings - Statistical Computing Section, Arizona State University

Webb, Wilson and Chong (2004), *An Analysis of Quasi-complete Binary Data with Logistic Models: Applications to Alcohol Abuse Data*, Journal of Data Science

Wedderburn, R. W. M. (1976), *On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models*, Biometrika, 63, 27-32

Software utilizado

Este trabalho foi realizado em SPSS 17.0 for Windows.